Bioinformatics tools for phylogeny and visualization

Yanbin Yin

Fall 2015

Homework assignment 5

1. Take the MAFFT alignment
    http://cys.bios.niu.edu/yyin/teach/PBB/
    purdue.cellwall.list.lignin.fa.aln as input and use MEGA5 to build a
    phylogenetic tree

2. Try maximum likelihood (ML), neighbor-joining (NJ) and maximum
    parsimony (MP) algorithms with 100 bootstrap replications and
    compare the running time and the topology of the resulting trees.
    If encounter errors, try to use the HELP link to find out and solve it

3. Color the branches and leafs in the resulting ML tree graph using
    different colors for different gene subfamilies

Homework assignment 5 Cont.

4. Export the tree as a newick format file

5. Prepare a color definition file for different gene subfamilies (see step 3); upload the newick tree file and the color definition file to iTOL to display the tree

Write a report (in word or ppt) to include all the operations and screen shots.

Office hour:

Due on Oct 20 (send by email)

Tue, Thu and Fri 2-4pm, MO325A

Or email: yyin@niu.edu

Outline

- Introduction to phylogenetic analysis

- Hands on practice of MEGA 5 and iTOL

Phylogenetics is the science of estimating the evolutionary past, in the case of molecular phylogeny, based on the comparison of DNA or protein sequences:

- Study the evolution of genomes and gene families (duplication and transfer)

- Study the diversity of genes or fragments

- Cluster homologous sequences into subfamilies based on evolutionary history

- Infer functions for unknown genes
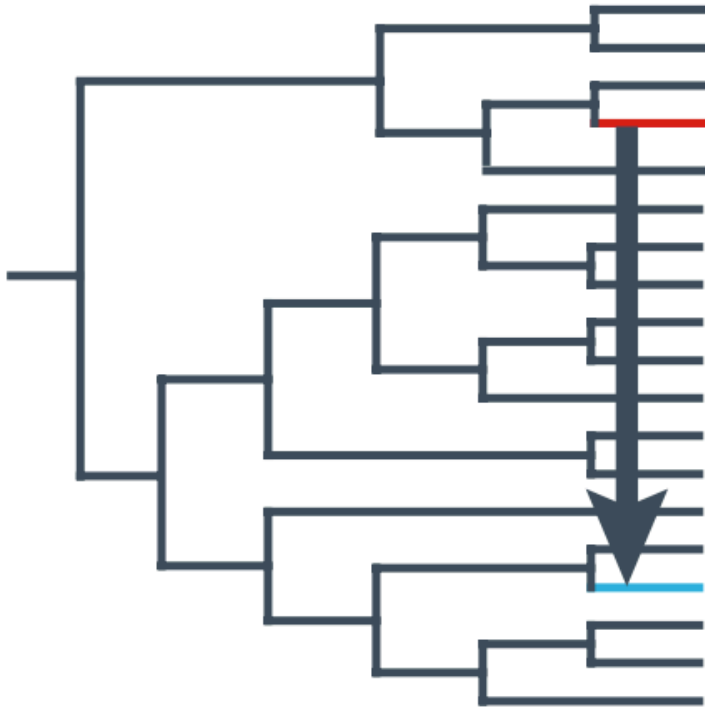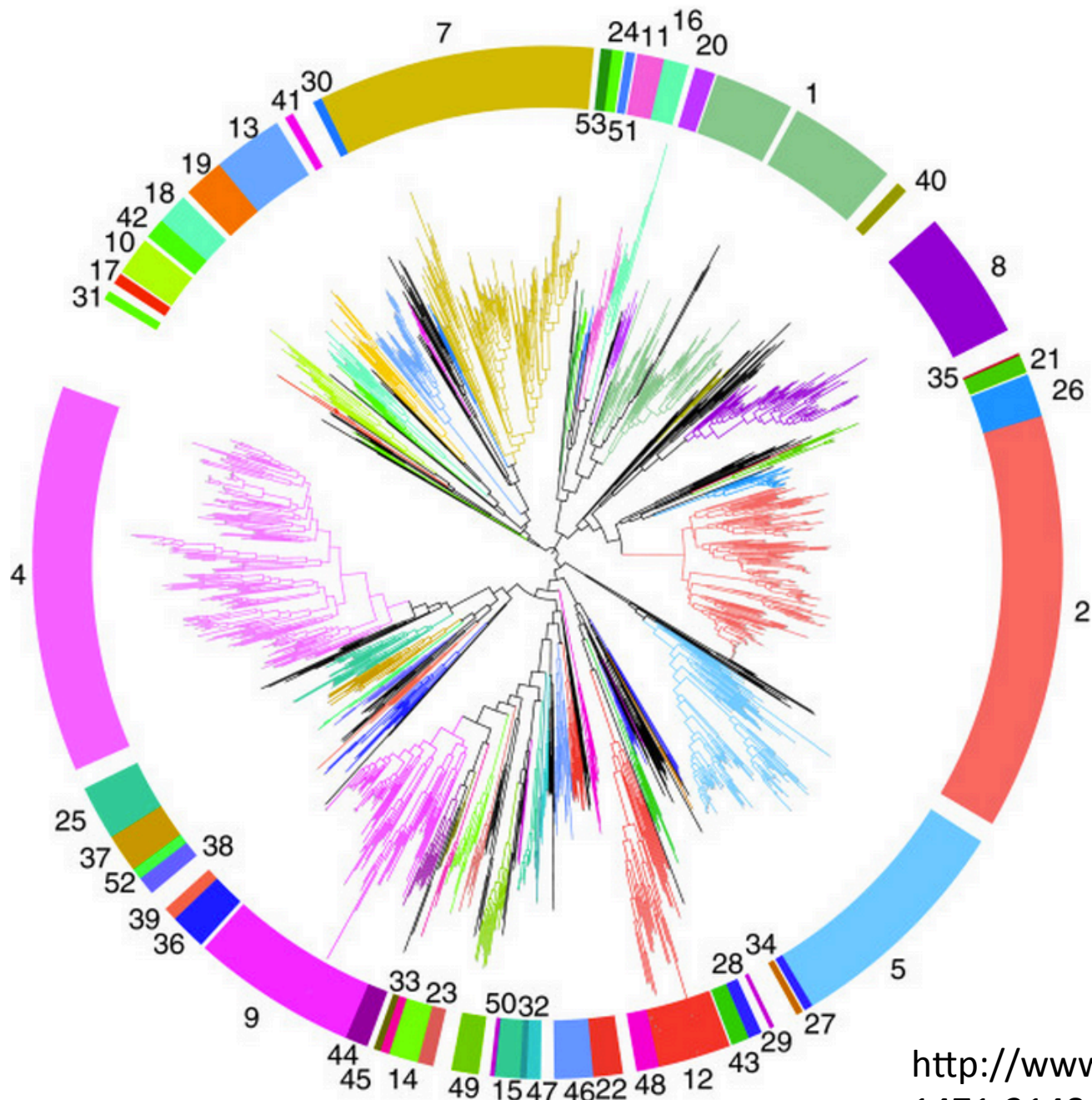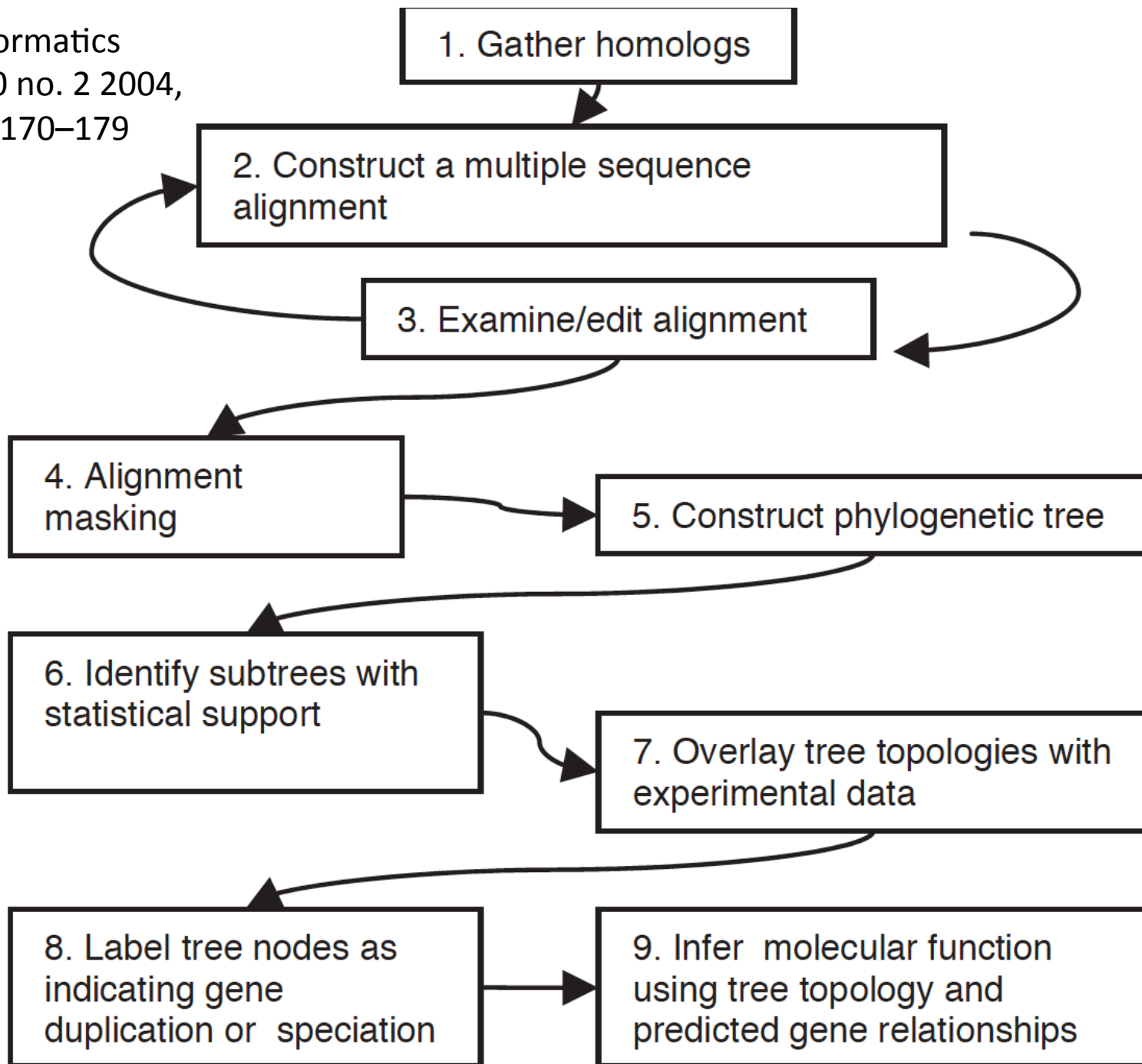
A simple case of horizontal gene transfer
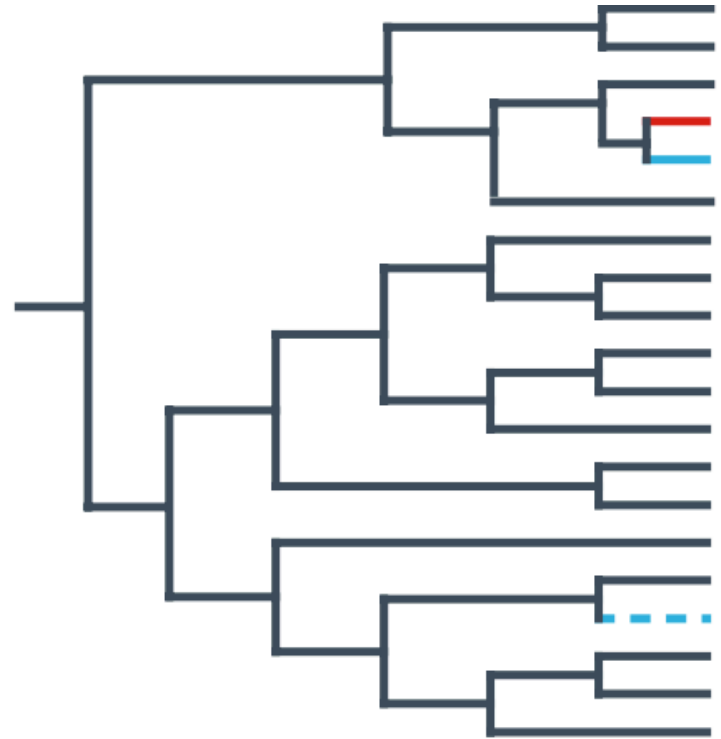
Figure 1.

Phylogenetic tree of family GH5. In this circular phylogram, the branches corresponding to subfamilies 1–53

7

1. Gather homologs

2. Construct a multiple sequence alignment

3. Examine/edit alignment

4. Alignment masking

5. Construct phylogenetic tree

6. Identify subtrees with statistical support

7. Overlay tree topologies with experimental data

8. Label tree nodes as indicating gene duplication or speciation

9. Infer molecular function using tree topology and predicted gene relationships

Step 1. Assembling a dataset
BLAST, FASTA, domain/family based (HMMER)

Step 2. Multiple sequence alignment
MAFFT, MUSCLE, Clustal Omega

Step 3. Phylogeny reconstruction
MEGA5, PHYML, RAxML, GARLI, MrBayes, FastTree

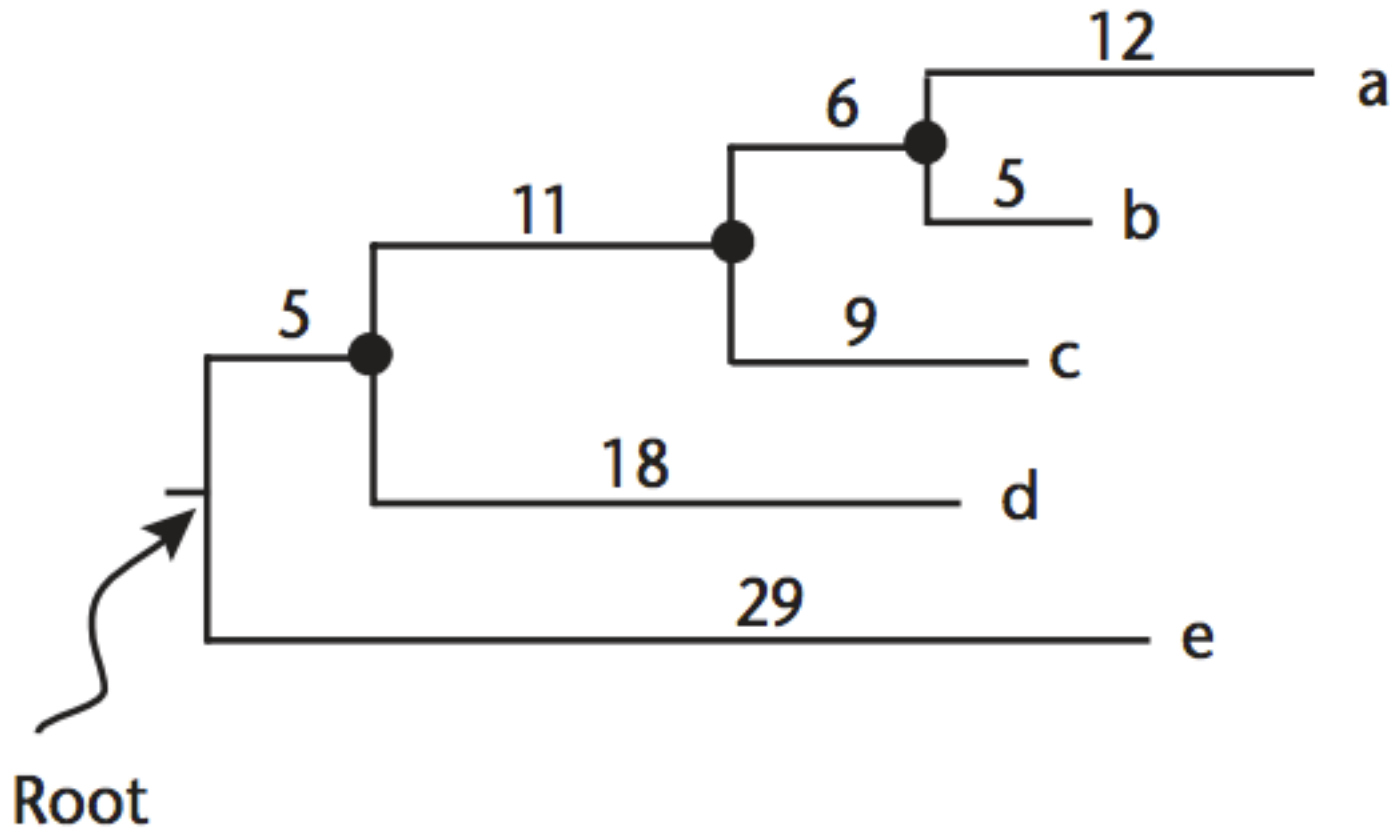## Step 4. Tree visualization
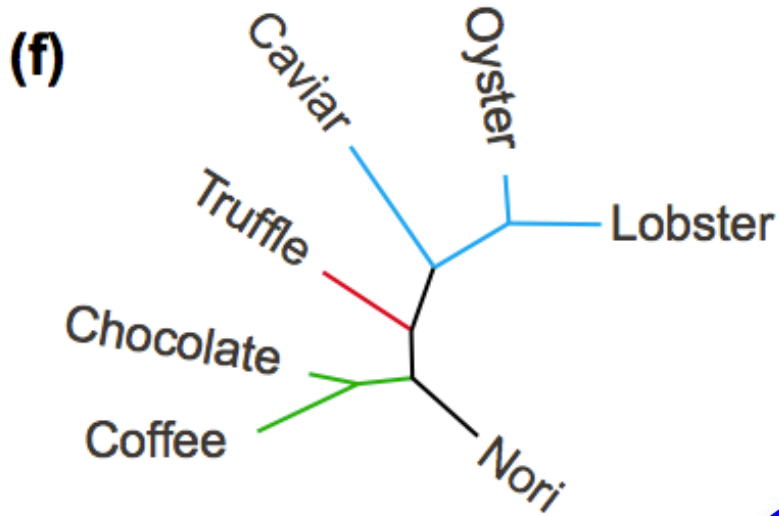TreeView, TreeDyn, MEGA5, iTOL

Unrooted tree

a

12

b

5

Branch length

c

9

6

11

Node

Internal node (inferred)

d

18

Exterior branch

34

Terminal node (actual seq)

Leaf node

Operational taxonomic unit (OTU)
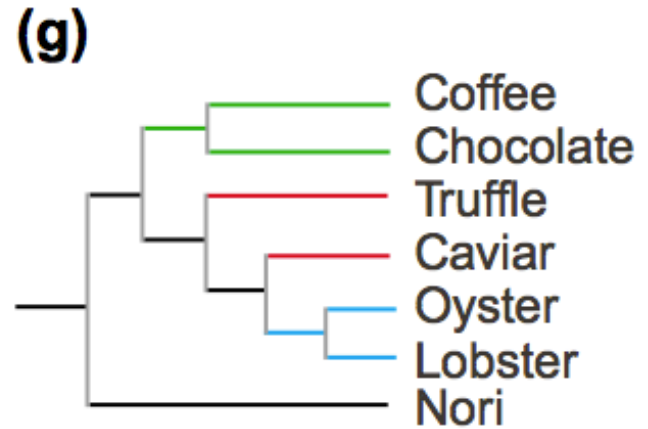
e
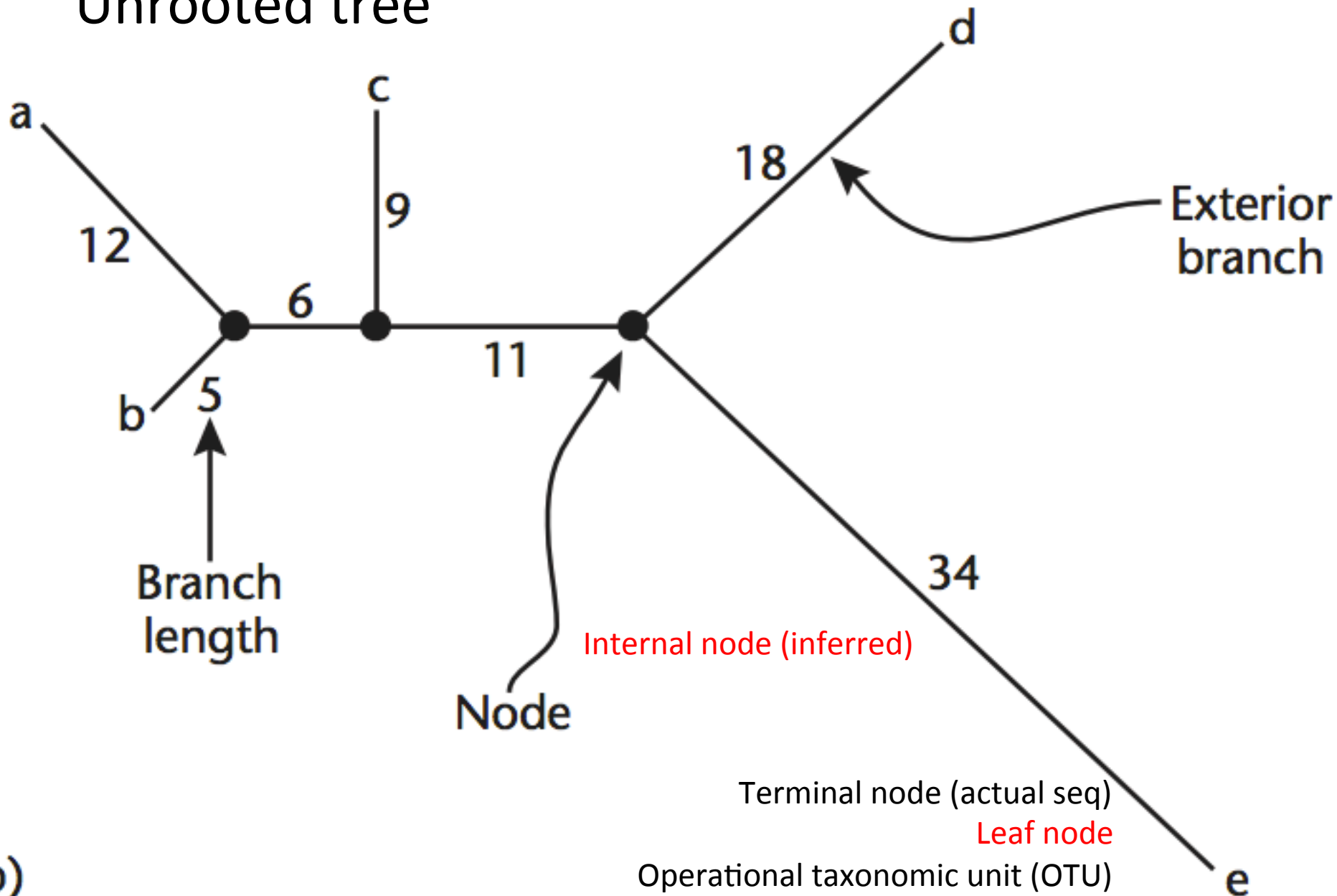
(b)

Rooted tree
Root is often selected based on prior knowledge

Branches are drawn with lengths proportional to the divergence (difference) between two nodes
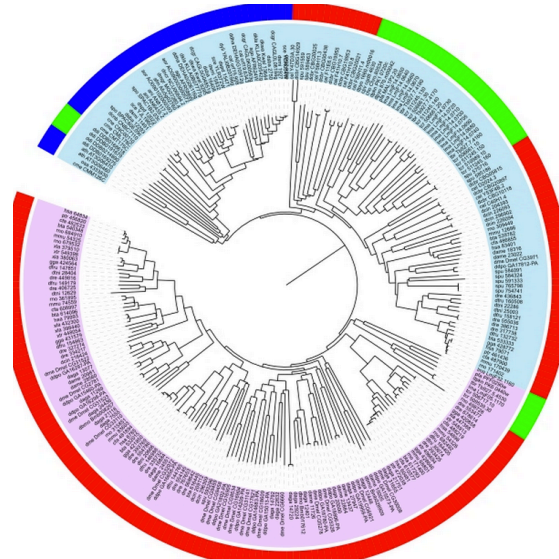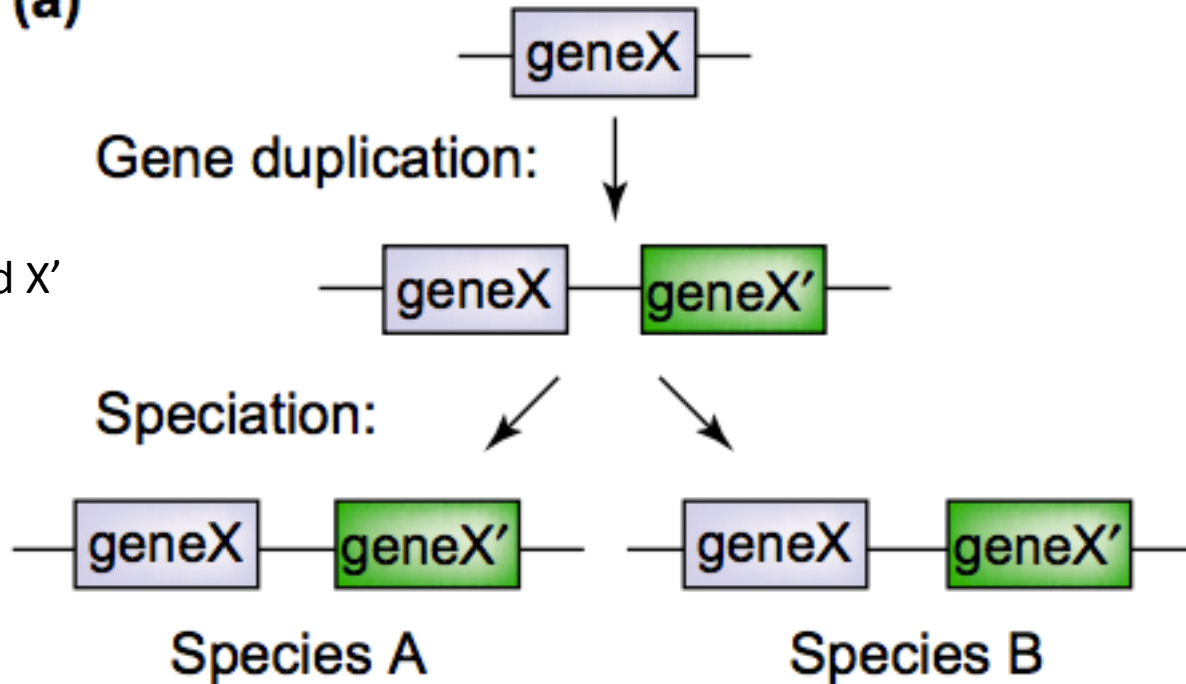
Radial view

Rectangular view

(f)

(g)

Coffee
Chocolate
Truffle
Caviar
Oyster
Lobster
Nori

TRENDS in Genetics

Circular view

12

(a)

geneX

Gene duplication:

Paralog: X and X'

geneX — geneX'

Speciation:

geneX — geneX'   geneX — geneX'

Species A   Species B

Ortholog:
X in A and X in B   What about X in A and X' in B?
X' in A and X' in B   They are called out-paralog (not often used)

All the four genes together are called an orthologous group

MEGA: Molecular Evolutionary Genetics Analysis

MEGA is an integrated tool for conducting sequence alignment, inferring phylogenetic trees, mining web-based databases, estimating rates of molecular evolution, inferring ancestral sequences, and testing evolutionary hypotheses. MEGA is used by biologists in a large number of laboratories for reconstructing the evolutionary histories of species and inferring the extent and nature of selective forces shaping the evolution of genes and species

Mega was developed as a software with GUI

The most cited phylogenetics analysis software package

MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0   [PDF] from researchgate.net
K Tamura, J Dudley, M Nei, S Kumar - Molecular biology and evolution, 2007 - SMBE
Abstract We announce the release of the fourth version of **MEGA software**, which expands
on the existing facilities for editing DNA sequence data from autosequencers, mining Web-
databases, performing automatic and manual sequence alignment, analyzing sequence ...
Cited by 23629   Related articles   All 11 versions   Cite   Saved

MEGA: a biologist-centric **software** for evolutionary analysis of DNA and protein sequences   [HTML] from oxfordjournals.org
S Kumar, M Nei, J Dudley, K Tamura - Briefings in bioinformatics, 2008 - Oxford Univ Press
Abstract The Molecular Evolutionary Genetics Analysis (**MEGA**) **software** is a desktop
application designed for comparative analysis of homologous gene sequences either from
multigene families or from different species with a special emphasis on inferring ...
Cited by 2035   Related articles   All 22 versions   Cite   Save

MEGA3: integrated **software** for molecular evolutionary genetics analysis and sequence   [PDF] from oxfordjournals.org
alignment
S Kumar, K Tamura, M Nei - Briefings in bioinformatics, 2004 - Oxford Univ Press
... From its inception, our goal for the **MEGA software** has been to make available a wide variety
of statistical and computational methods for comparative sequence analysis in a user-friendly
environment.8–10 The first version of **MEGA**,10 released in 1993, was distributed to over ...
Cited by 12270   Related articles   All 24 versions   Cite   Save

MEGA: molecular evolutionary genetics analysis **software** for microcomputers
S Kumar, K Tamura, M Nei - Computer applications in the ..., 1994 - Oxford Univ Press
Abstract A computer program package called **MEGA** has been developed for estimating
evolutionary distances, reconstructing phylogenetic trees and computing basic statistical
quantities from molecular data. It is written in C++ and is intended to be used on IBM and ...
Cited by 1760   Related articles   All 5 versions   Cite   Save

MEGA6: molecular evolutionary genetics analysis version 6.0   [PDF] from kumarlab.net
K Tamura, G Stecher, D Peterson, A Filipski... - Molecular biology and ..., 2013 - SMBE
... Abstract. We announce the release of an advanced version of the Molecular Evolutionary Genetics
Analysis (**MEGA**) **software**, which currently contains facilities for building sequence alignments,
inferring phylogenetic histories, and conducting molecular evolutionary analysis. ...
Cited by 498   Related articles   All 11 versions   Cite   Save

MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary   [HTML] from oxfordjournals.org
distance, and maximum parsimony methods
K Tamura, D Peterson, N Peterson, G Stecher... - Molecular biology and ..., 2011 - SMBE
... We also thank Mr Paul Billing-Ross for his help with computer simulations and Ms
Carol Williams for editorial support. The **MEGA software** project is supported by
research grants from National Institutes of Health to SK and MN. ...
Cited by 15969   Related articles   All 27 versions   Cite   Save

MEGA2: molecular evolutionary genetics analysis **software**   [PDF] from oxfordjournals.org
S Kumar, K Tamura, IB Jakobsen, M Nei - Bioinformatics, 2001 - Oxford Univ Press
... J. Mol. Evol., 41, 1152–1159. Kumar,S., Tamura,K. and Nei,M. (1994) **MEGA**: Molecular Evolu-
tionary Genetics Analysis **software** for microcomputers. Comput. Appl. Biosci., 10, 189–191.
Nei,M. and Kumar,S. (2000) Molecular Evolution and Phylogenet- ics. ...
Cited by 8850   Related articles   All 16 versions   Cite   Save

http://www.megasoftware.net/

Free download for different OSs, e.g. WINDOWS

it's free, but you need to fill out an on-line form to download

MEGA 6 (Windows)

MEGA is provided FREE for use in research and education (see terms below).

To download MEGA version 6 (release notes), please fill in the information requested below about how and where MEGA is used. This anonymous information is important for obtaining funding support for the maintenance and further development of MEGA.

Important Notice: By downloading MEGA, you are agreeing not to redistribute the MEGA software and associated materials partially or fully in any form.

This version is for Windows only. It is not intended to be used on Mac or Linux systems, as it has not been optimized for those platforms.

United States   Country*

I am a ... (check one or more)*
- ☐ Researcher
- ☐ Instructor
- ☐ Student
- ☐ Professor
- ☐ Other

MEGA5 is already installed on MO444 computers

find MEGA5 in the start->program->MEGA

What kind of institution is Comcast Cable ? (check one or more)*
- ☐ University/College
- ☐ Research Institute
- ☐ Industry
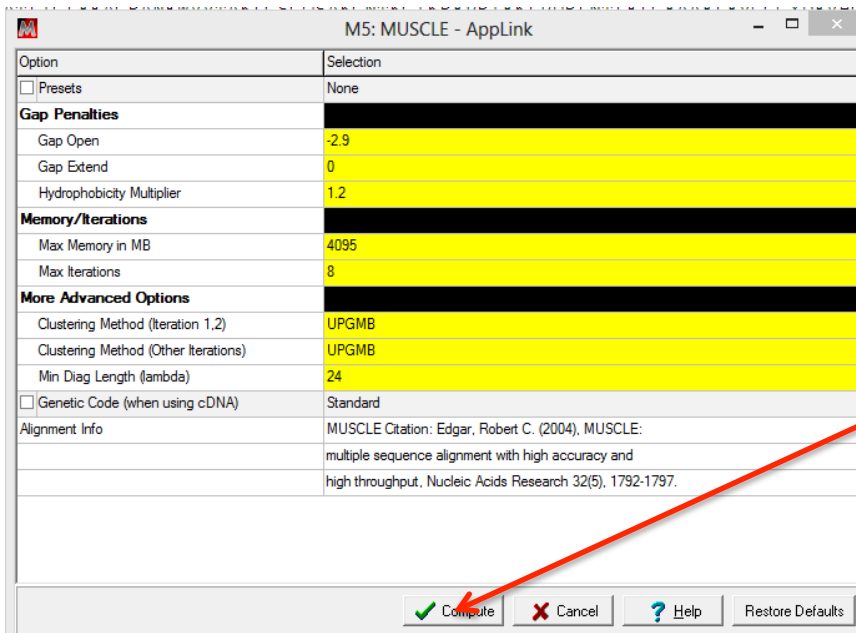- ☐ Government
- ☐ Other

Download MEGA   Reset

We're gonna use MEGA to do the alignment first, then build the phylogeny

Click on Open a File, then copy paste the URL
http://cys.bios.niu.edu/yyin/teach/PBB2013/cesa-pr.fa

Align the seq first

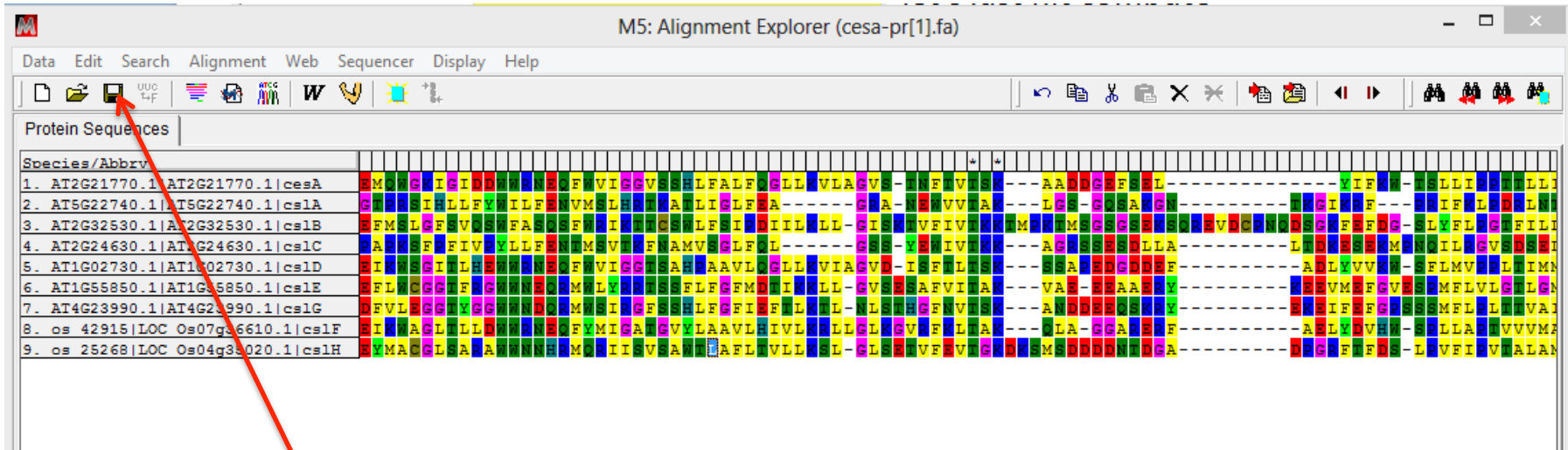Click on alignment then choose align by muscle
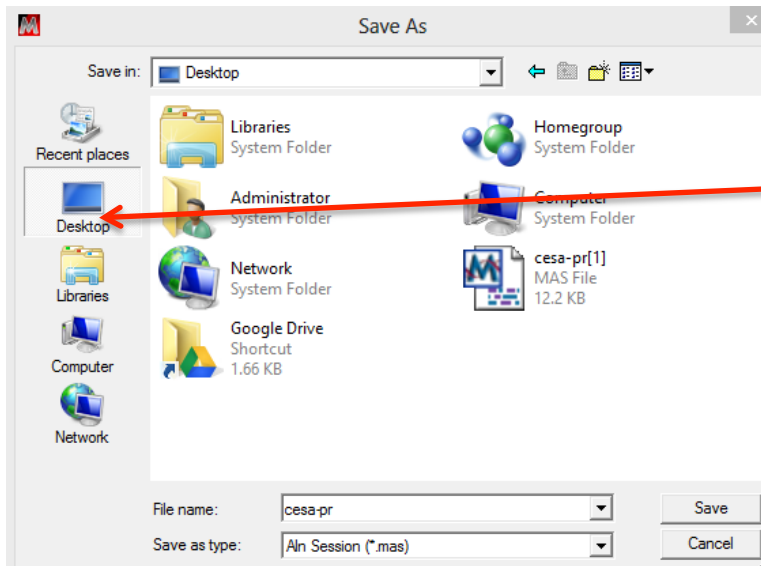
The alignment explorer popped out

Select yes

Popped out window to allow option change
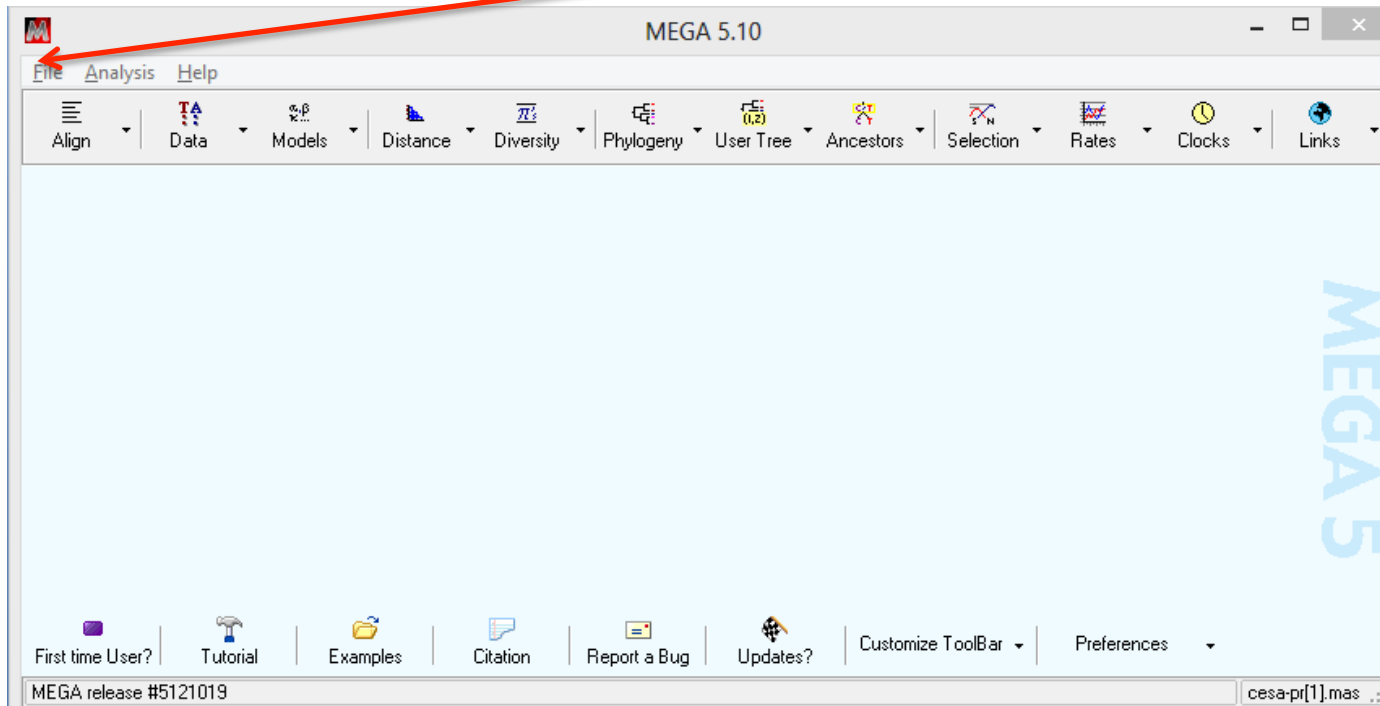
let's just hit compute

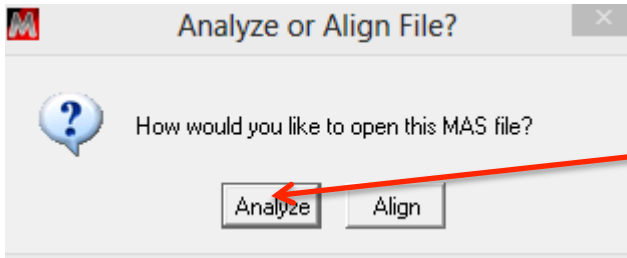Now the alignment explorer shows the aligned seqs
Next hit the save icon to save the alignment as a MEGA format

Now I saved it in the desktop folder

Now go back to the main window, click on File to open the saved mas file

This time choose analyze as it's an aligned file

This window changed, meaning the data is loaded; we can build the tree now

You may choose from a list of different building algorithms basically, maximum likelihood is the most accurate but also the slowest neighbor-joining and maximum parsimony are also very popular and faster if you have over 50 sequences or longer sequences
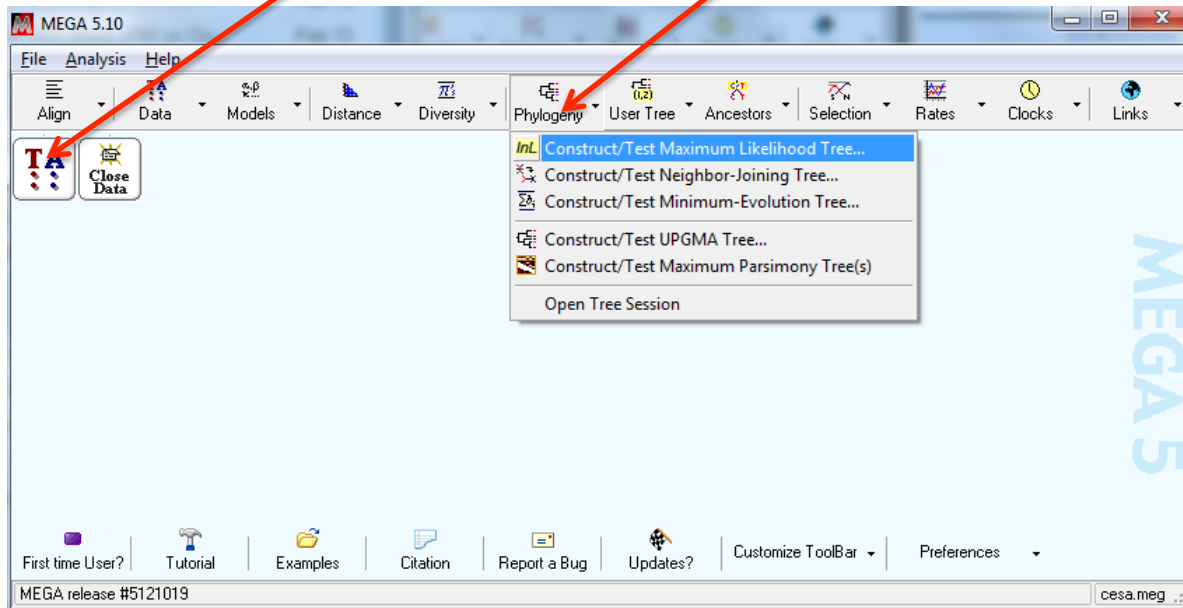
Phylogenetic trees are calculated by applying mathematical models to infer evolutionary relationships between molecules or organisms (here sequences), based on a set of characters that describe their differences.

Four main categories of phylogenetic reconstruction methods:

1. Maximum parsimony approaches create trees using the minimum number of ancestors needed to explain the observed characters

2. Distance matrix methods, such as neighbor joining, allow more sophisticated evolutionary models than parsimony

3. Maximum likelihood methods search a set of tree and evolutionary models to find the ones most likely to generate the observed characters

4. Bayesian approaches offer more flexibility, as they allow optimization of all aspects of a tree (model, topology, branch length)
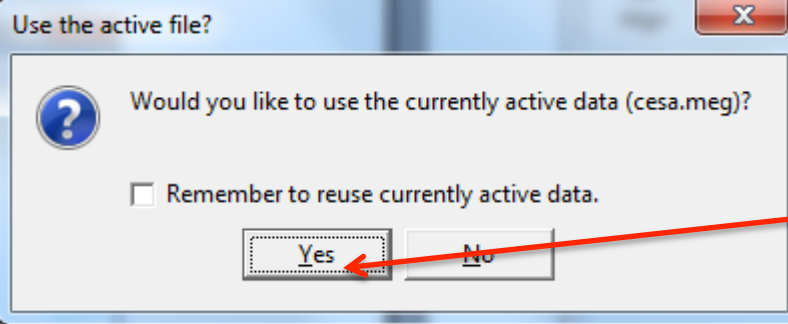
Maximum likelihood and Bayesian, in general, outperformed neighbor joining and maximum parsimony in terms of tree reconstruction accuracy.

In general, our results indicate that as alignment error increases, topological accuracy decreases.
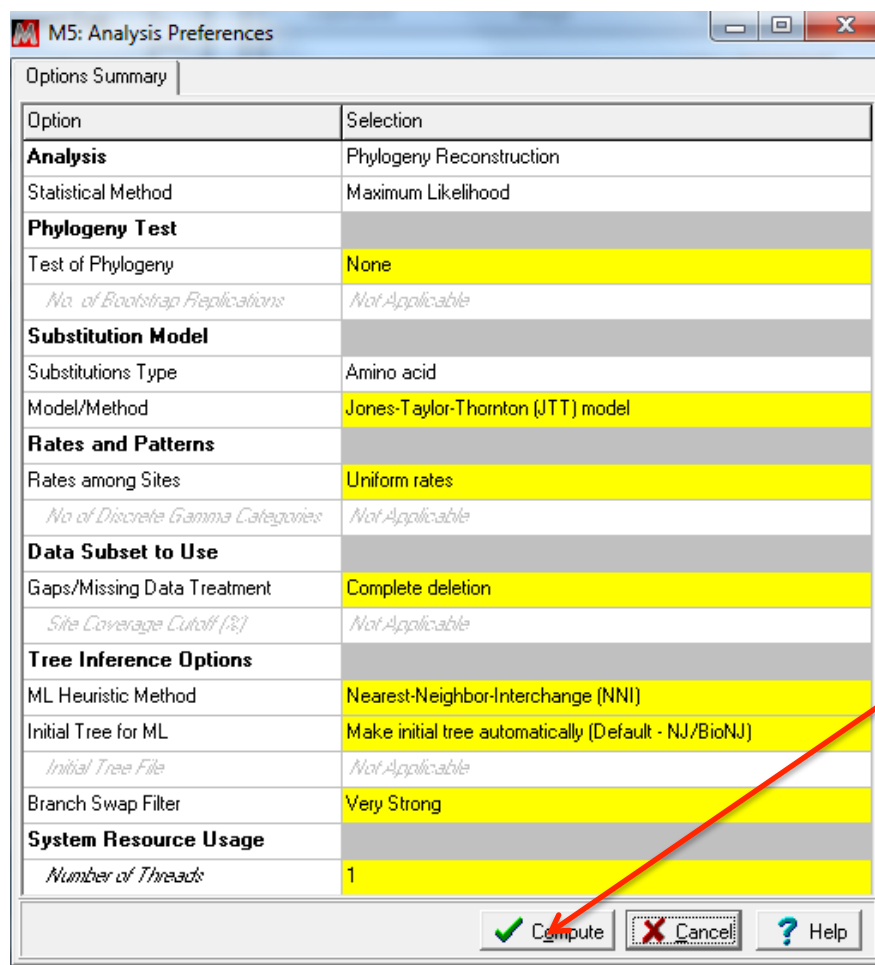
Results also indicated that as the length of the branch and of the neighboring branches increase, alignment accuracy decreases, and the length of the neighboring branches is the major factor in topological accuracy.

Over the variety of conditions tested, Bayesian trees estimated from DNA sequences that had been aligned according to the alignment of the corresponding protein sequences were the most accurate, followed by Maximum Likelihood trees estimated from DNA sequences and Parsimony trees estimated from protein sequences

Choose yes

You may choose parameters for tree building

Let's just hit compute

the tree graph is shown after it's done
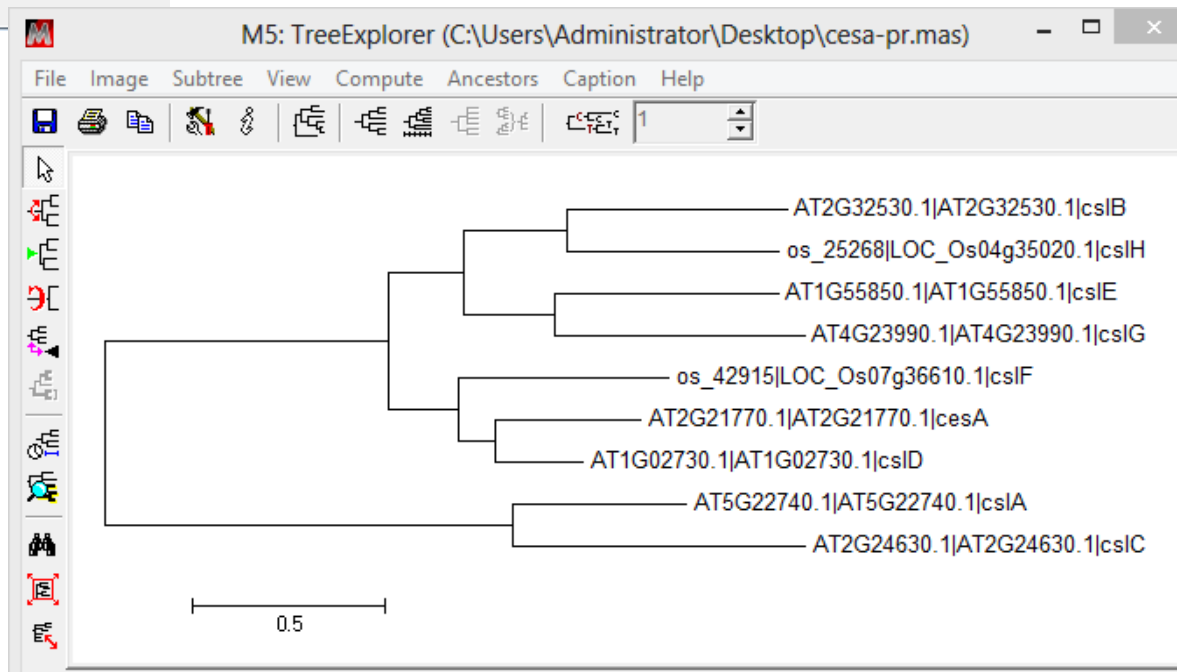
if we want to have statistical values on the clustering

this time we want to choose neighbor-joining algorithm because it is much faster than maximum likelihood.
Here we also want to choose bootstrap method to test the phylogeny then we will have statistical values for
each node.

Option	Selection
Analysis	Phylogeny Reconstruction
Scope	All Selected Taxa
Statistical Method	Neighbor-joining
Phylogeny Test	
Test of Phylogeny	Bootstrap method
No. of Bootstrap Replications	100
Substitution Model	
Substitutions Type	Amino acid
Model/Method	Jones-Taylor-Thornton (JTT) model
Rates and Patterns	
Rates among Sites	Uniform rates
Gamma Parameter	*Not Applicable*
Pattern among Lineages	Same (Homogeneous)
Data Subset to Use	
Gaps/Missing Data Treatment	Pairwise deletion
Site Coverage Cutoff (%)	*Not Applicable*

M5: Analysis Preferences — Options Summary
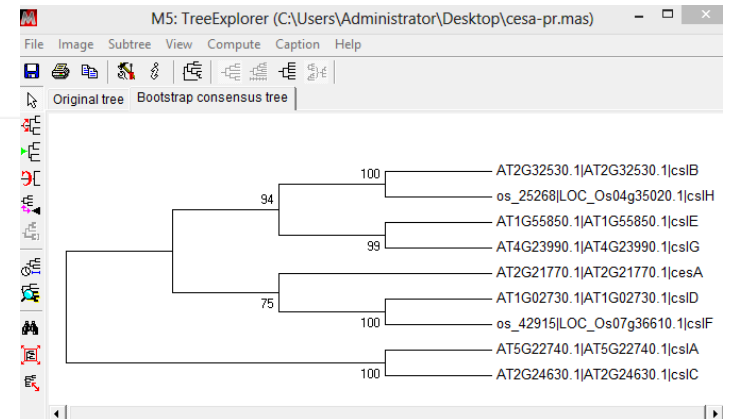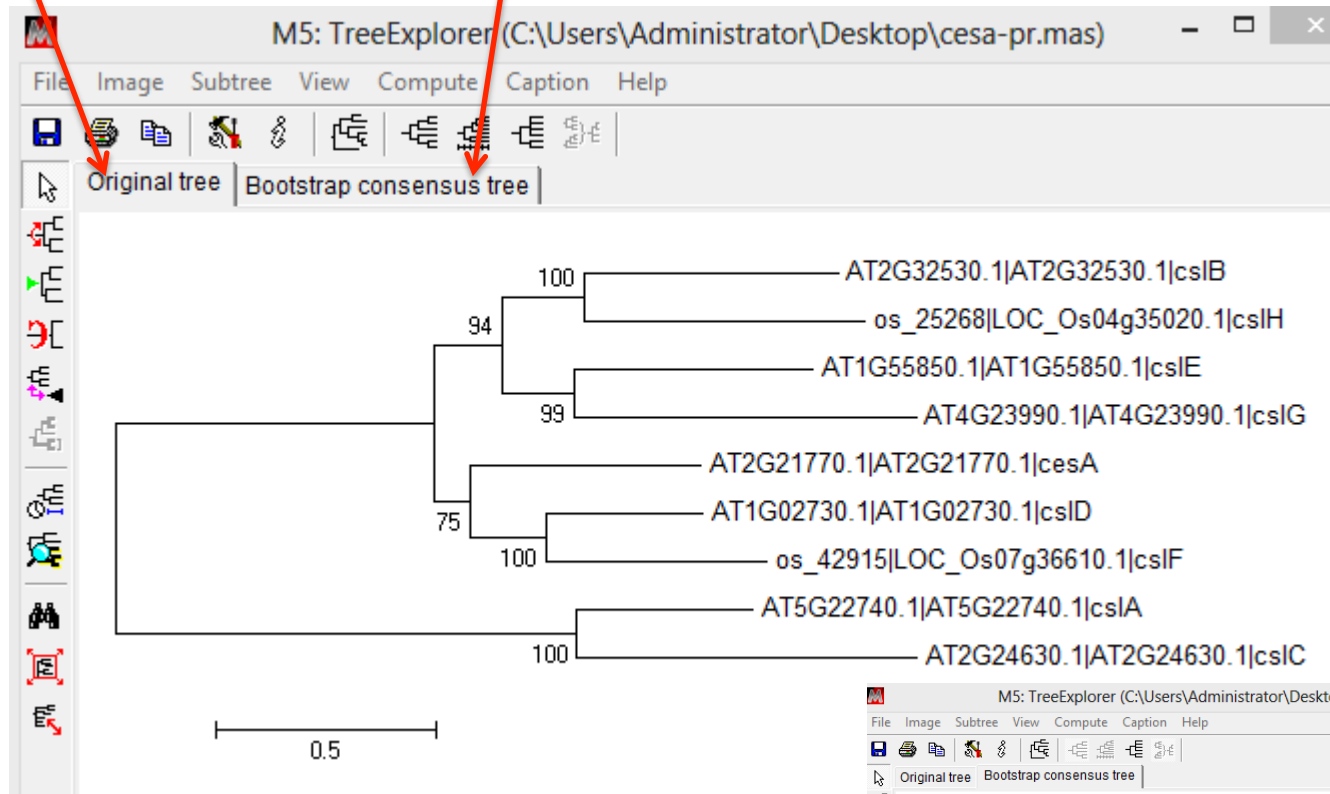
✔ Compute   ✘ Cancel   ? Help

Now change here

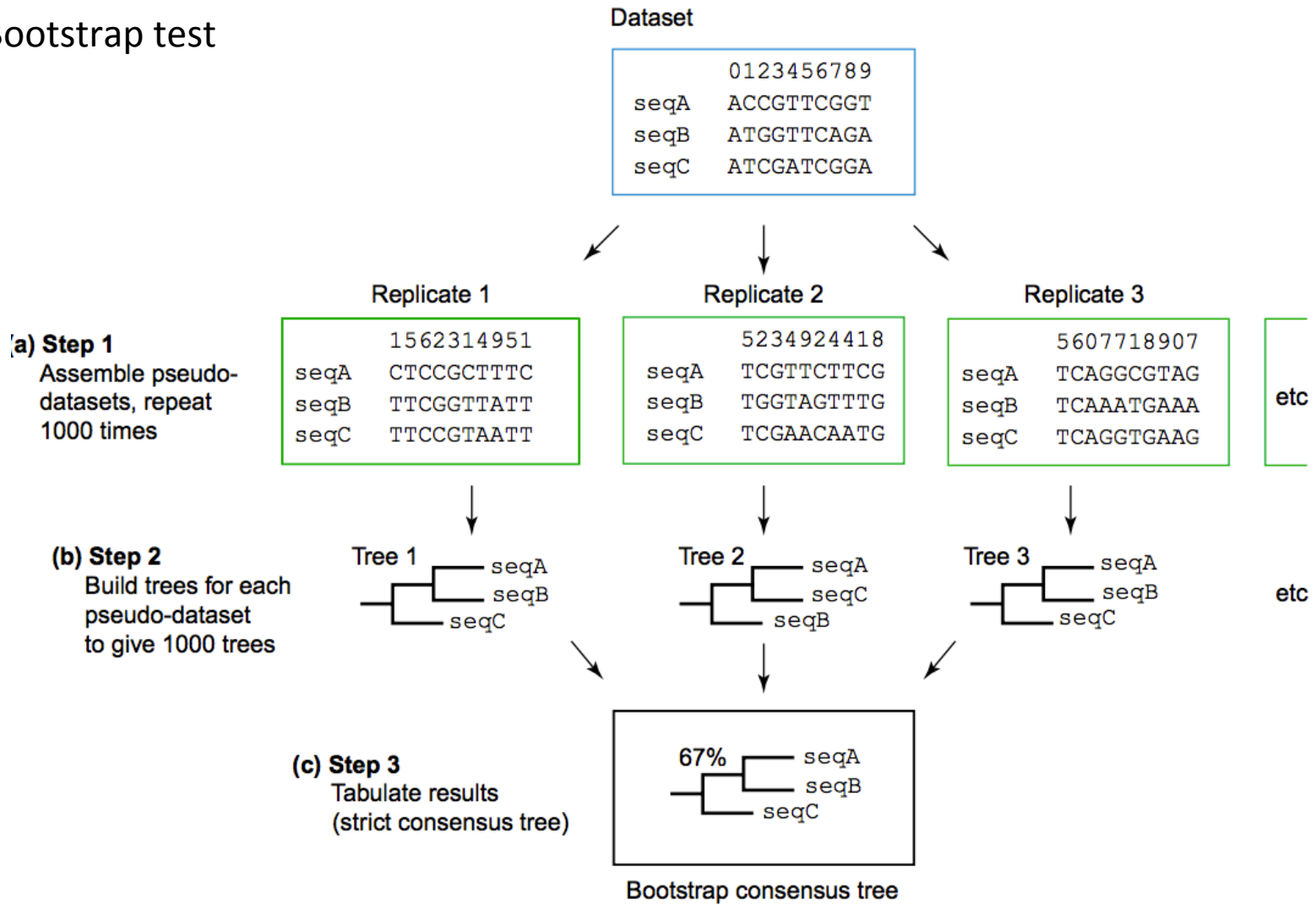Yellow are where you can change

To learn what do these options mean
Click on help

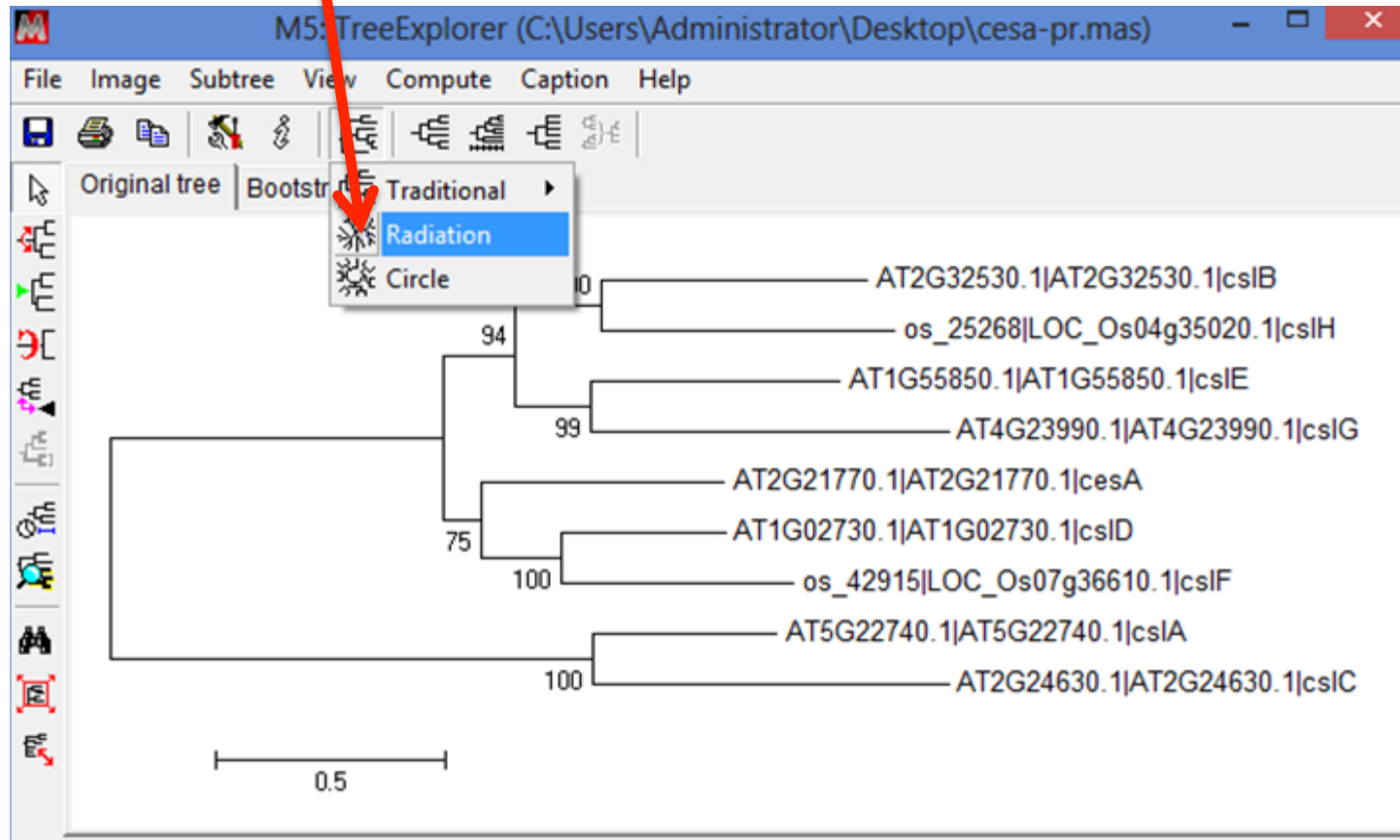This is the original tree with bootstrap support values at each internal node
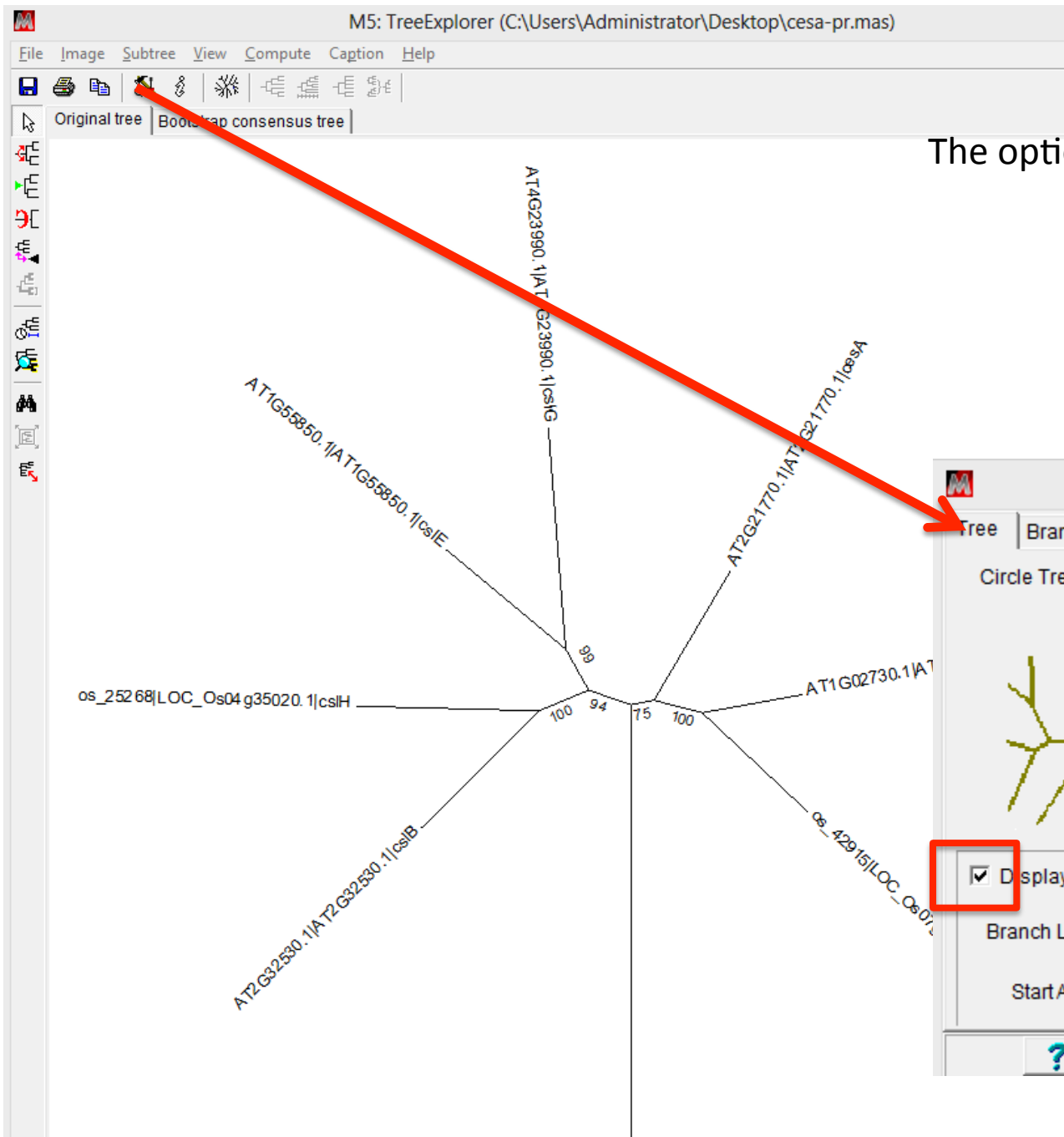
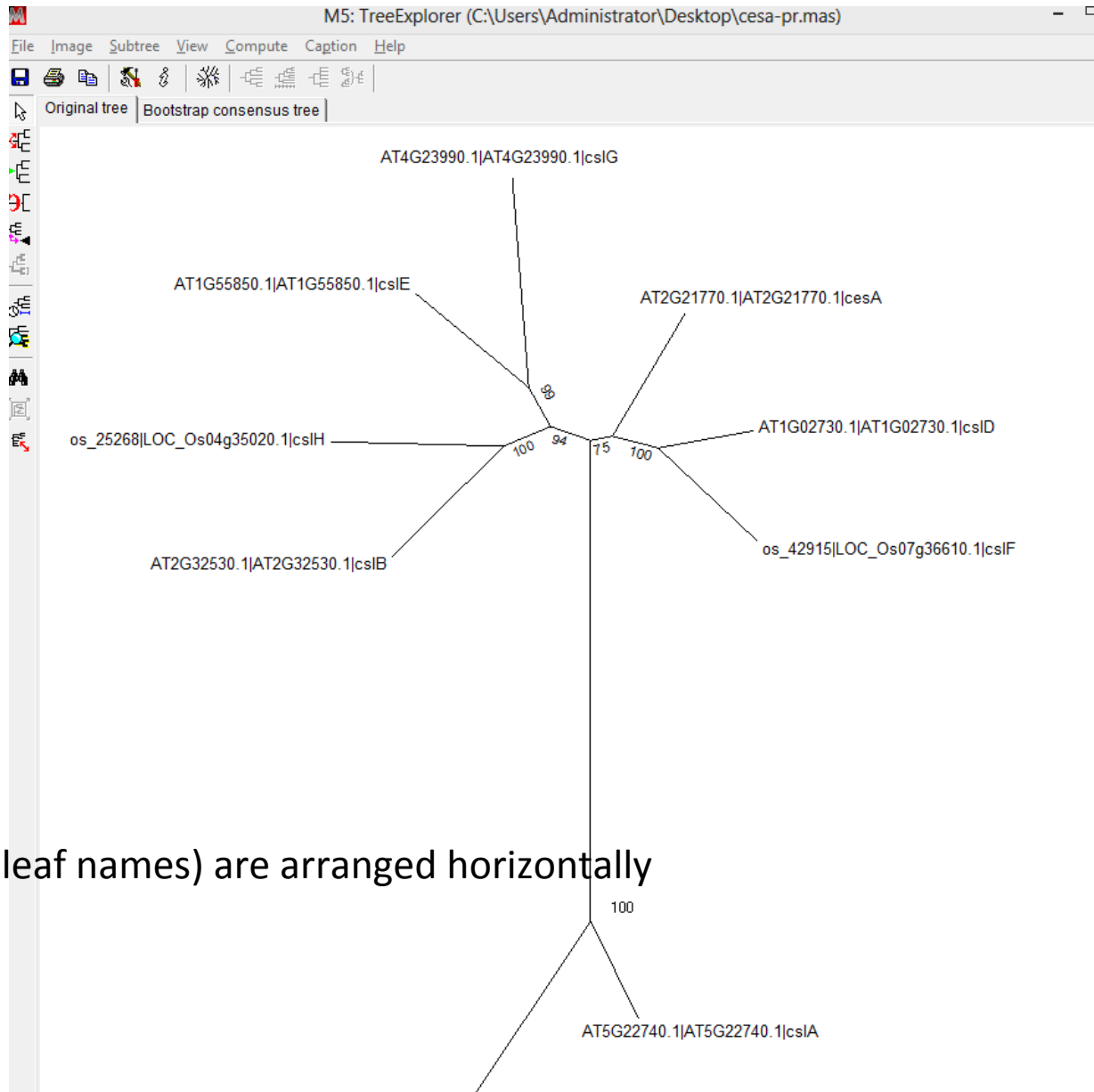Consensus tree from bootstrap test

Bootstrap test

TRENDS in Genetics

Different presentation views of phylograms
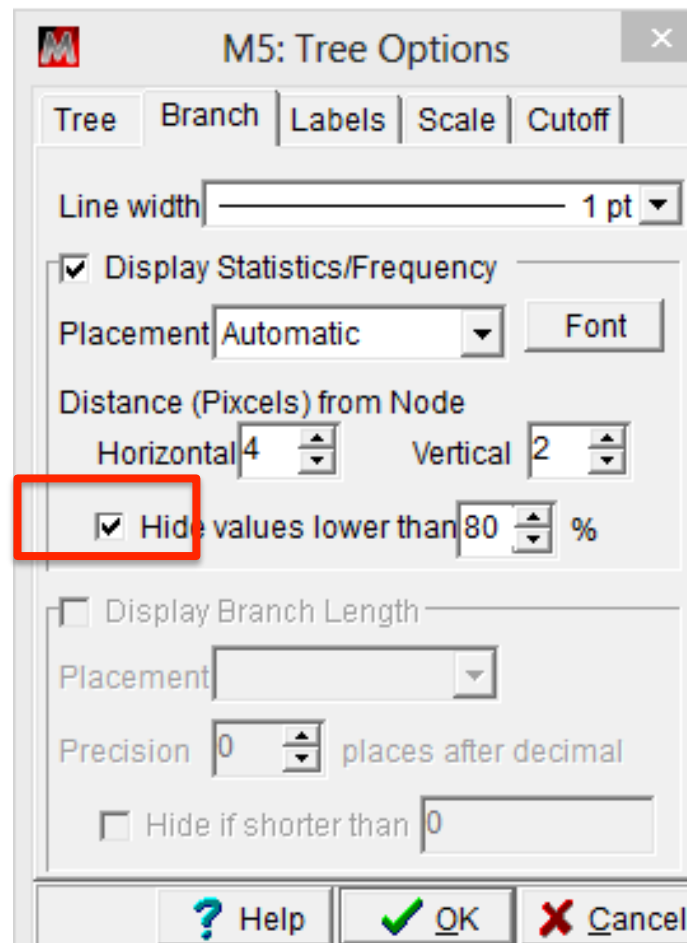
The option window

31

Now the IDs (leaf names) are arranged horizontally

To only show good bootstrap values higher than certain values

Export phylogram as image file,
Click Image then save as

M5: TreeExplorer (C:\Users\Administrator\Desktop\cesa-pr.mas)

File   Image   Subtree   View   Compute   Caption   Help

Save Current Session   Ctrl+S
Export Current Tree (Newick)
Export Current Tree (Clock)
Export All Trees (Newick)

Write Tree in a Table format

Export Group Names
Import Group Names

Show Information   Ctrl+I

Print
Print in a sheet
Printer setup...

Exit Tree Explorer   Ctrl+Q

AT4G23990.1|AT4G23990.1|cslG

AT1G55850.1|AT1G55850.1|cslE

AT2G21770.1|AT2G21770.1|cesA

os_25248|LOC_Os04g35020.1|cslH

AT1G02730.1|AT1G02730.1|cslD

100   94   100

os_42915|LOC_Os07g36610.1|cslF

AT2G32530.1|AT2G32530.1|cslB

Export the text format file that
defines phylogeny topology
File then Export Newick file

35

Open the saved newick format file in notepad

((((AT2G32530.1|AT2G32530.1|cslB:0.57646262,'os_25268|LOC_Os04g35020.1|cslH':
0.63658065)1.0000:0.18712502,(AT1G55850.1|AT1G55850.1|cslE:
0.54168375,AT4G23990.1|AT4G23990.1|cslG:
0.77646829)0.9900:0.16421052)0.9400:0.15649299,(AT2G21770.1|AT2G21770.1|cesA:
0.52631255,(AT1G02730.1|AT1G02730.1|cslD:0.35504124,'os_42915|
LOC_Os07g36610.1|cslF':
0.50349483)1.0000:0.17352695)0.7500:0.08201111)1.0000:0.72454177,(AT5G22740.1|
AT5G22740.1|cslA:0.39871493,AT2G24630.1|AT2G24630.1|cslC:
0.77203016)1.0000:1.04968340);

Not for human read!!!

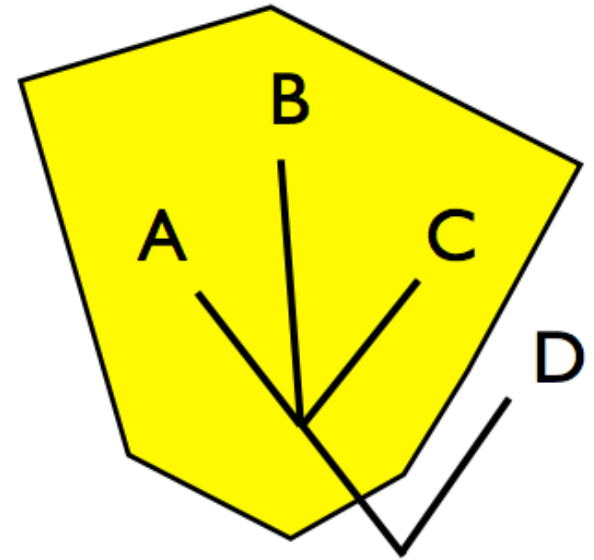Newick format uses parenthesis to group two nodes at a time to describe the groupings
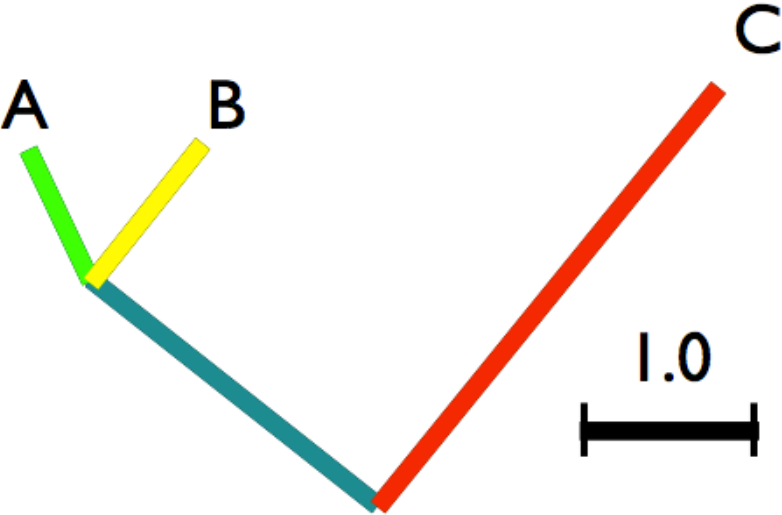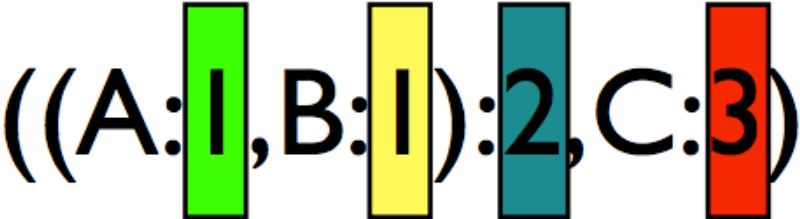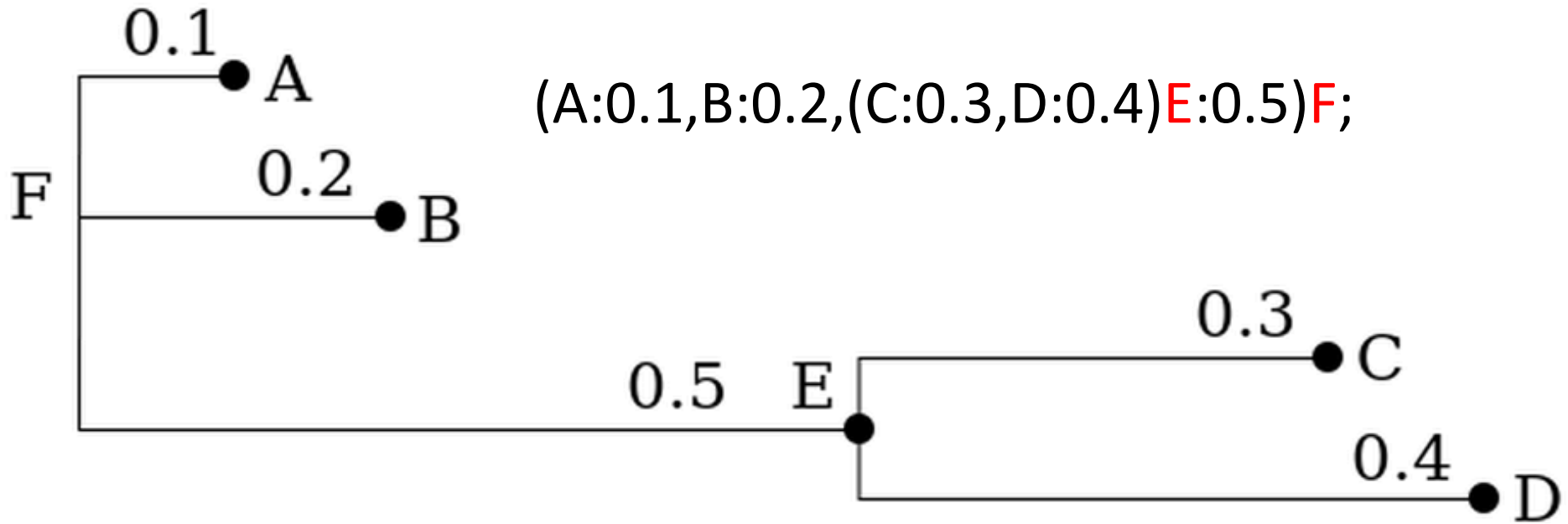
A most simplified example
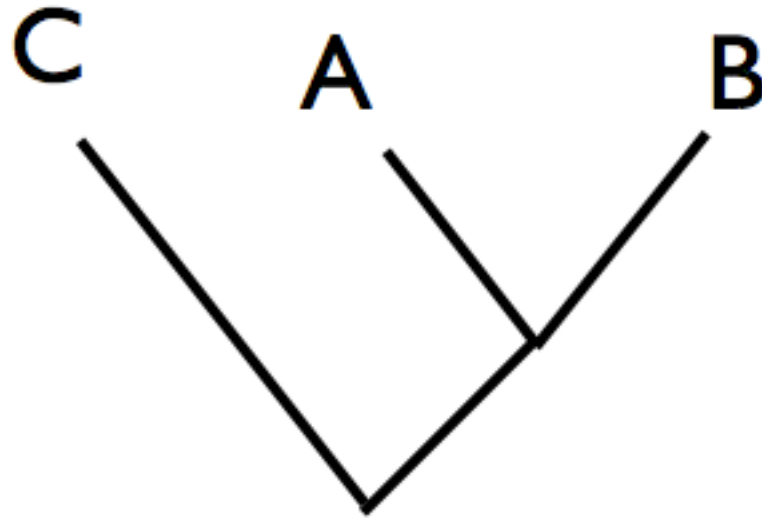
$$((A,B),C);$$

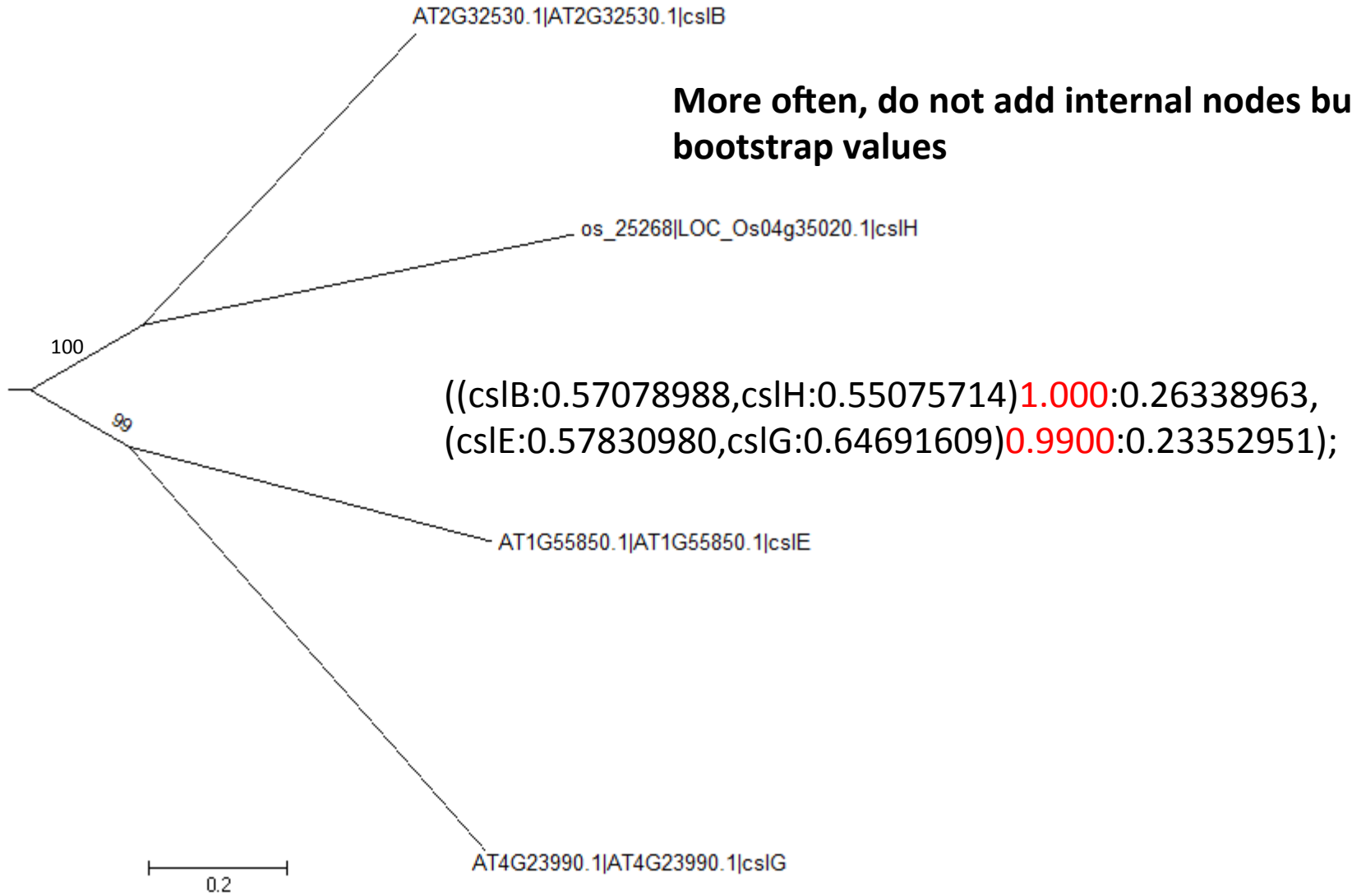polytomy/multifurcation

$((A,B,C),D);$

Add the branch length

((A:1,B:1):2,C:3)

Add the internal node name

(A:0.1,B:0.2,(C:0.3,D:0.4)E:0.5)F;

E and F and inferred nodes, not from the input

AT2G32530.1|AT2G32530.1|cslB

os_25268|LOC_Os04g35020.1|cslH

More often, do not add internal nodes but add bootstrap values

100

99

((cslB:0.57078988,cslH:0.55075714)1.000:0.26338963,
(cslE:0.57830980,cslG:0.64691609)0.9900:0.23352951);

AT1G55850.1|AT1G55850.1|cslE

0.2

AT4G23990.1|AT4G23990.1|cslG

41
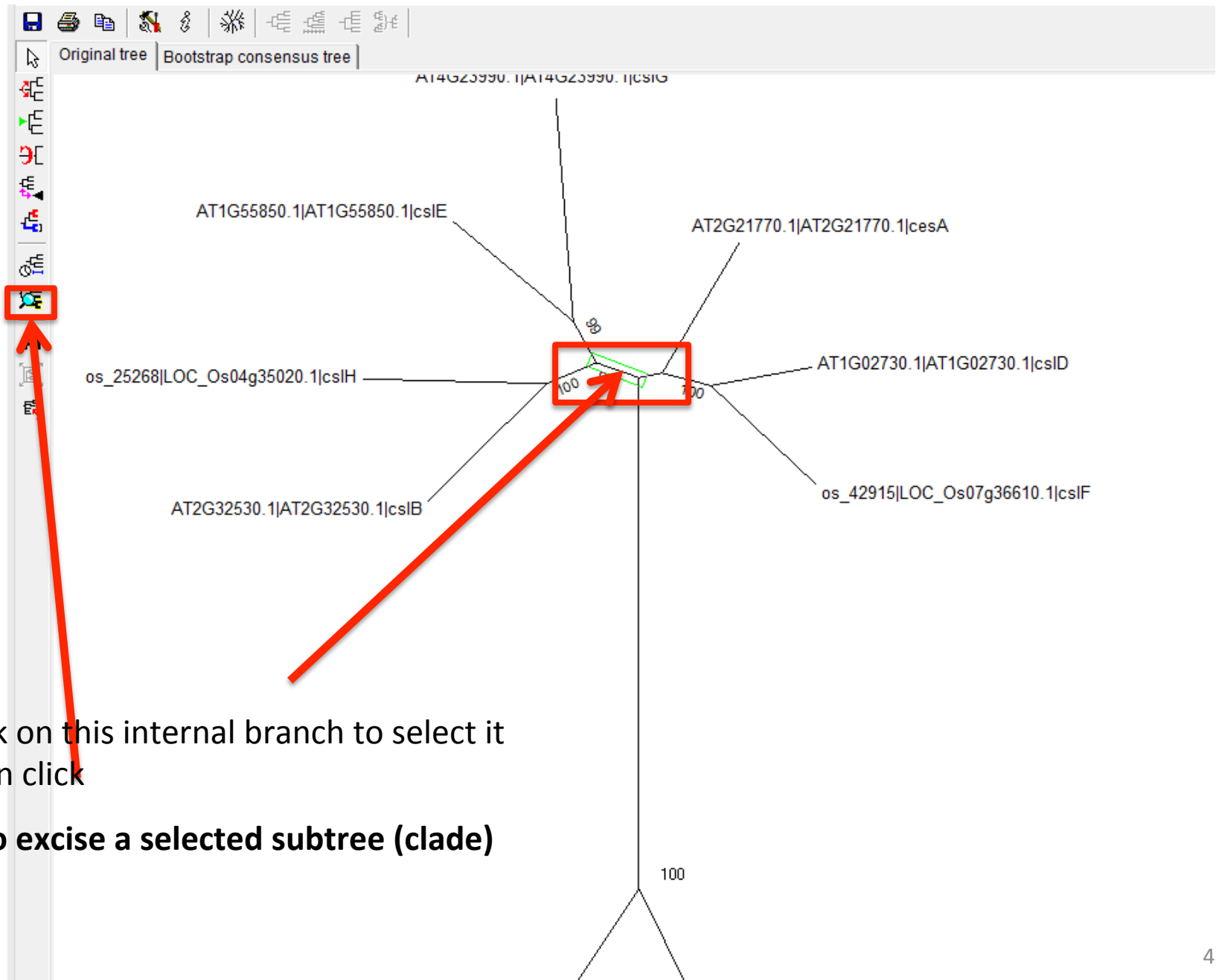
Click on this internal branch to select it
Then click

To excise a selected subtree (clade)

To color branches

AT4G23990.1|AT4G23990.1|cslG

AT1G55850.1|AT1G55850.1|cslE

AT2G21770.1|AT2G21770.1|cesA

Os04g35020.1|cslH

AT1G02730.1|AT1G02730.1|cslD

32530.1|AT2G32530.1|cslB

os_42915|LOC_Os07g36610.1|cslF

Right click on the internal branch

M5: Subtree Drawing O...

Property | Display | Image

Name/Caption

[|] Font

Node/Subtree Marker

Shape [None ▼]   Color [███ ▼]

☐ Apply to Taxon Markers

Branch Line

Lines [⌐═ ▼]   Color [███ ▼]

Width [████████████ 5 pt ▼]

Style [────────── Solid ▼]

☐ Overwrite Downstream   ☐ Default

? Help   ✓ OK   ✗ Cancel

Swap
Flip
Compress/Expand
Place Root
Divergence Time
Selected Subtree

100

AT5G22740.1|AT5G22740.1|cslA

43
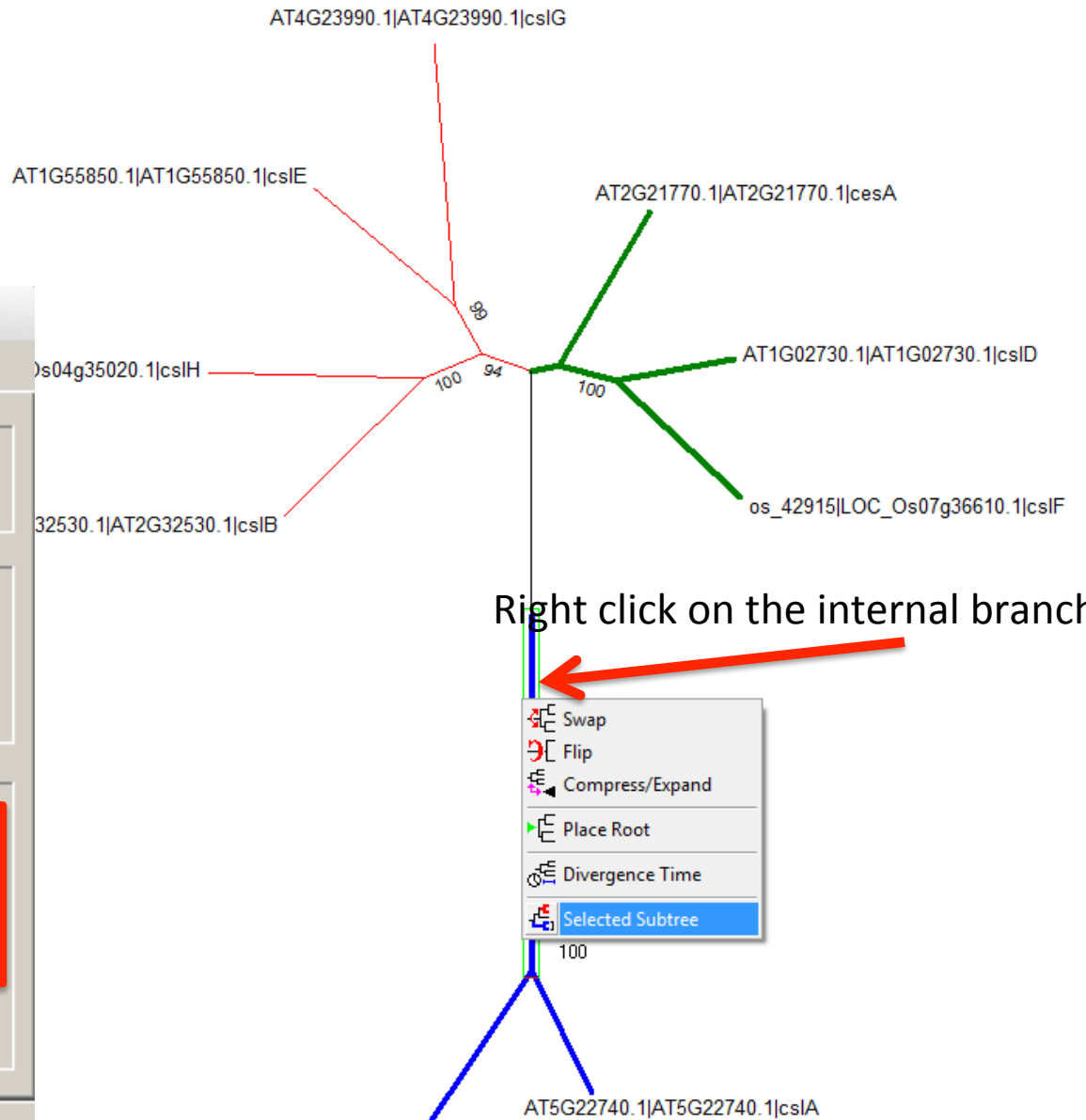
30.1|AT2G32530.1|cslB

Change the fonts of leaf names

AT2G24630.1|AT2G24630.1|cslC

100

Swap
Flip
Compress/Expand

Place Root

Divergence Time

Selected Subtree

M5: Subtree Drawing O...

Property | Display | Image

☑ Display Caption   ☐ Align Vertically

☑ Display Bracket

Style] Square ▼   Color ████ ▼

Line Width 1   Pixels

☑ Display Taxon Names   Font

☑ Display Node Markers

☑ Display Taxon Markers

☐ Compress Subtree

Vertical Unit 2   Pixels/Taxon

Fill Pattern ████ ▼

☐ Overwrite Downstream   ☐ Default

? Help   ✓ OK   ✗ Cancel

Font

Font: Font Style: Size:
Arial   Regular   10   OK

Arial   Regular   10   Cancel
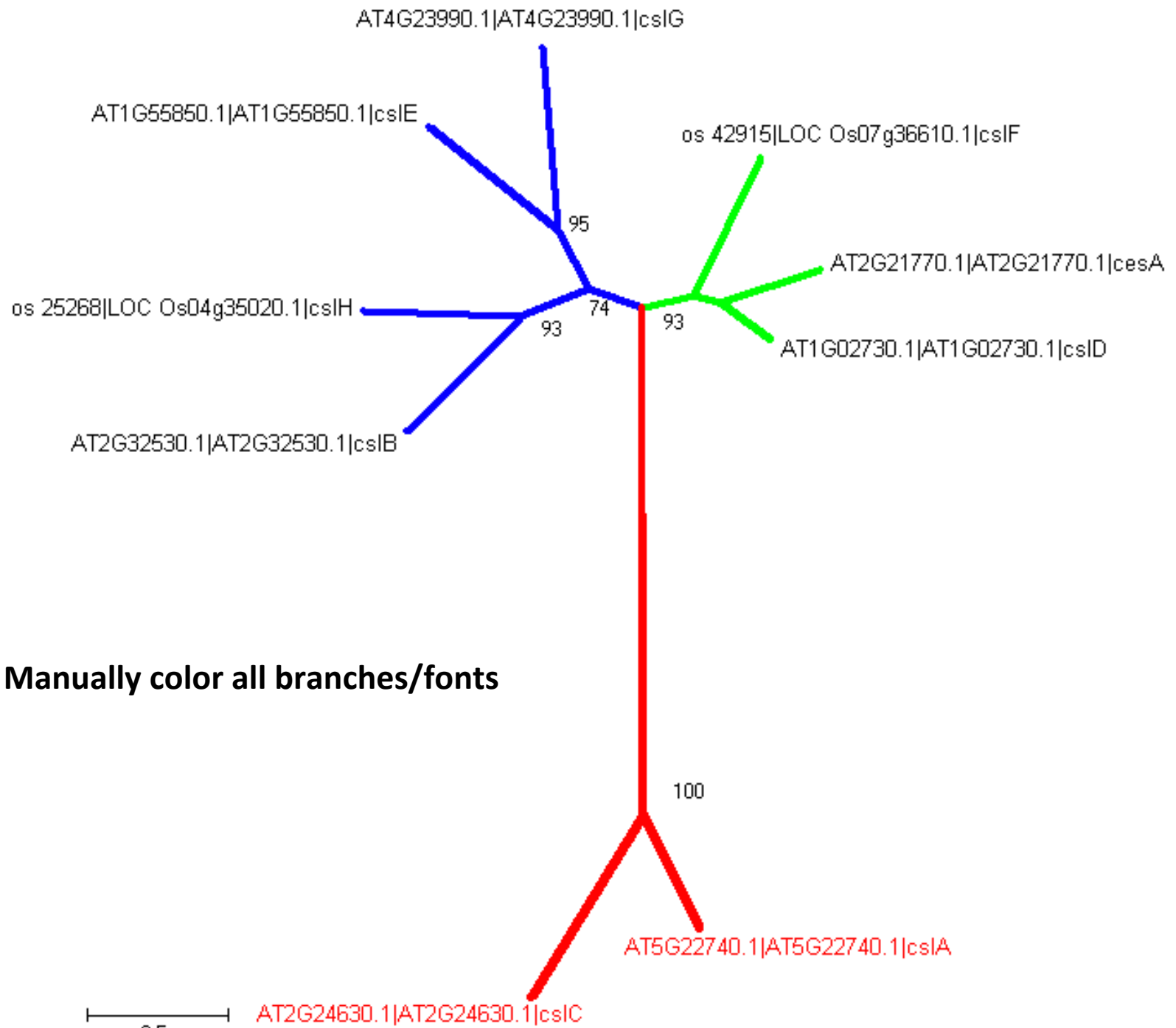Arial Black   Italic   11
Arial Narrow   Bold   12
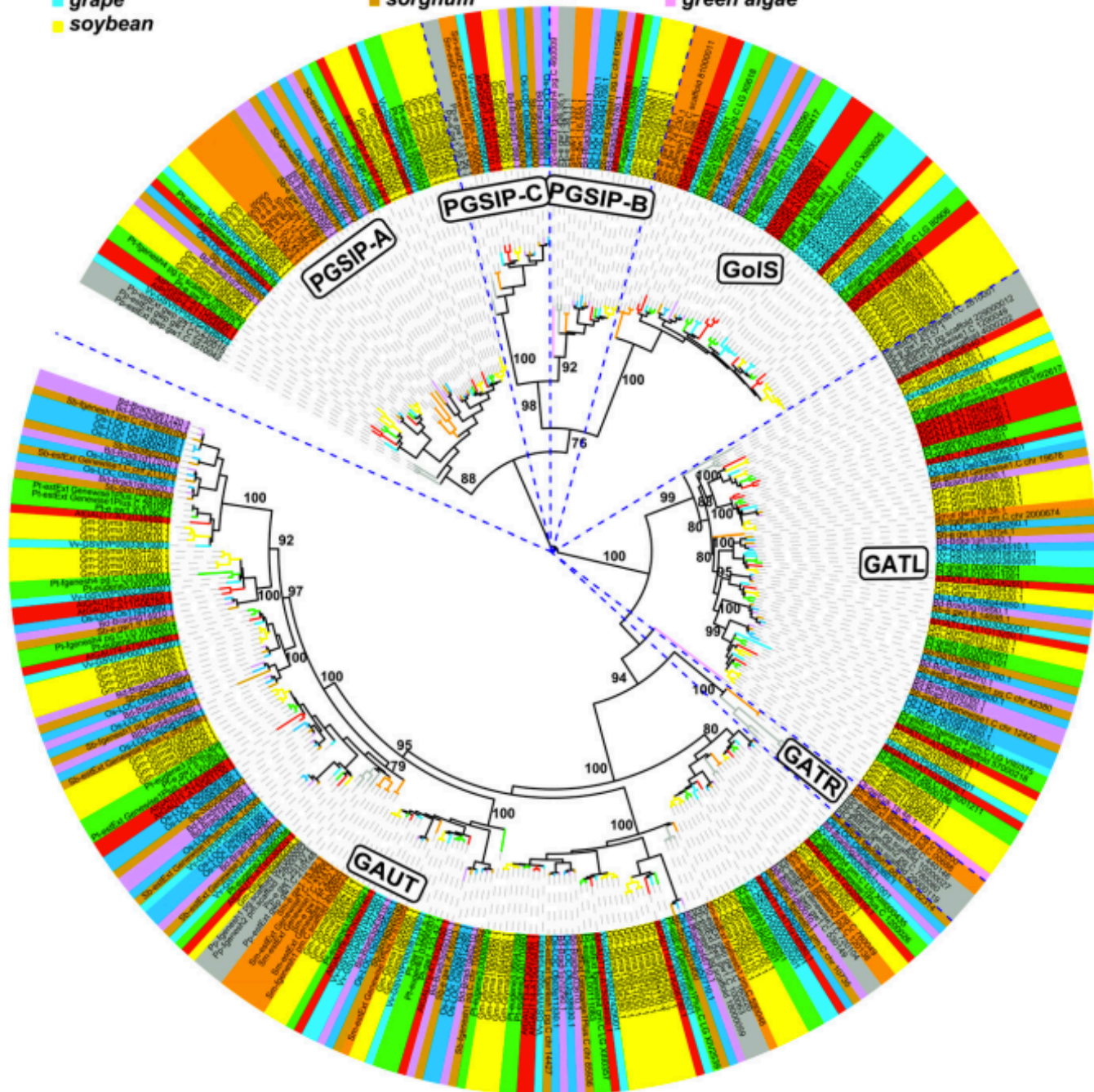Arial Rounded MT E ▼   Bold Italic ▼   14 ▼

Effects   Sample
☐ Strikeout
☐ Underline   AaBbYyZz

Color: Script:
Red ▼   Western ▼

Manually color all branches/fonts

What if we have hundreds of genes?

http://itol.embl.de/

Automatically define branch colors by uploading a **color definition file**

You can define your own colors for each branch/leaf separately. Use standard hexadecimal color notation (for example, #ff0000 for red)
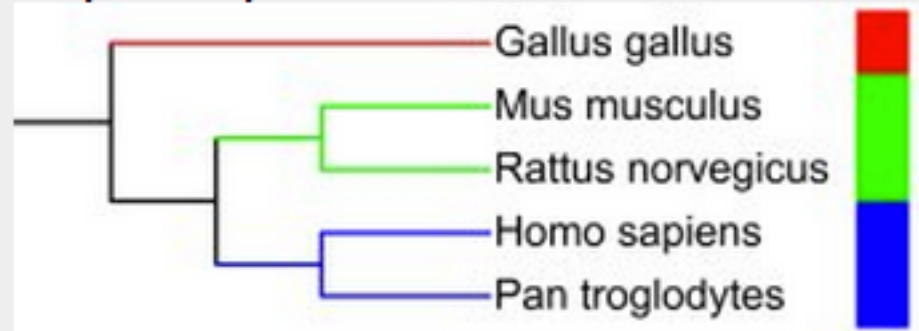
http://www.w3schools.com/html/html_colors.asp

Example:

Dataset file:

```
Gallus_gallus,#ff0000
Rattus_norvegicus,#00ff00
Mus_musculus,#00ff00
Homo_sapiens,#0000ff
Pan_troglodytes,#0000ff
```

Graphical representation:

http://itol.embl.de/help/help.shtml

http://cys.bios.niu.edu/yyin/teach/PBB/cesa-pr.fa.col

```
AT2G21770.1|AT2G21770.1|cesA       #00ff00
AT5G22740.1|AT5G22740.1|cslA       #ff0000
AT2G32530.1|AT2G32530.1|cslB       #0000ff
AT2G24630.1|AT2G24630.1|cslC       #ff0000
AT1G02730.1|AT1G02730.1|cslD       #00ff00
AT1G55850.1|AT1G55850.1|cslE       #0000ff
AT4G23990.1|AT4G23990.1|cslG       #0000ff
os_42915|LOC_Os07g36610.1|cslF     #00ff00
os_25268|LOC_Os04g35020.1|cslH     #0000ff
```

((((AT2G32530.1|AT2G32530.1|cslB:0.57078988,os_25268|LOC_Os04g35020.1|cslH:
0.55075714)0.9300:0.26338963,(AT1G55850.1|AT1G55850.1|cslE:
0.57830980,AT4G23990.1|AT4G23990.1|cslG:
0.64691609)0.9500:0.23352951)0.7400:0.19857786,(os_42915|LOC_Os07g36610.1|
cslF:0.54191868,(AT2G21770.1|AT2G21770.1|cesA:0.37516472,AT1G02730.1|
AT1G02730.1|cslD:
0.22502015)0.6600:0.09521396)0.9300:0.18369951)1.0000:0.73286595,(AT5G22740.1|
AT5G22740.1|cslA:0.44848889,AT2G24630.1|AT2G24630.1|cslC:
0.75671710)1.0000:1.05517231);

| HOME | TREE OF LIFE | OTHER TREES | SHARED PROJECTS | DATA UPLOAD | SAVED VIEWS | HELP | ABOUT |

Upload a prune list for our tree of life

You can upload a list of NCBI taxonomy IDs and the tree of life will be pruned to include only the species from your file. There should one species I
download.

File with tax IDs: [Choose File] No file chosen   [Prune tree]

Upload your own tree

NEW! If you are using iTOL to upload your own trees, try creating a personal account. More info about the iTOL personal accounts is availab

Use this form to upload your own phylogenetic tree. It should be in plain text, in one of supported formats (Newick, Nexus or PhyloXML) . Please ch
features available.

You can either paste your tree into the box, or upload a file using the file selector below. Don't forget to select the correct tree format.

Paste or type the tree:	Upload a file which contains your tree:
Paste your tree here...	[Choose File] No file chosen **Tree format:** Newick ⬍ Make sure the file is plain text, and contains only trees in the selected format.

Optional information	
Tree name: [ ]	*if you don't specify a name, a numeric ID will be used instead*

Advanced options (show)

Upload   (if you're uploading extra data with your tree, fill the dataset section below before clicking 'Upload')

Upload color definition file

Upload datasets for your tree

iTOL can annotate phylogenetic trees with several types of data. Please check our help pages for th

| **Dataset 1** | Dataset 2 | Dataset 3 | Dataset 4 | Dataset 5 | Dataset |

Dataset 1 file:	Choose File cesa–pr.fa.col	please use plain text files only
Display label:	color	label will be used in the legend
Field delimiter:	◯ Space ⦿ Tab ◯ Comma	make sure the correct delimite.
Data type:	Color Strips ⇕	check the help pages for detai
Prevent overlap:	◯ Yes ⦿ No	if yes, other dataset types will s
Strip width:	25 pixels	strip will have the specified wid
Branch coloring:	◯ none (show only leaf boxes) ⦿ color branches and show boxes ◯ color branches only	should the branches be colore
Dataset 1 color:	#FF0000	used in the legends and for da

Upload   **Clear all fields**

ITOL INTERACTIVE TREE OF LIFE

Tree uploaded succesfully

Tree has 17 nodes in total, of which 9 are leaves.

Parsing Dataset 1 (color)

Succesfully parsed 9 entries in the Dataset 1.

Our internal ID for this tree is 13115621113447513620026080.

Use this ID on the data upload page to retrieve it later.

What now?

- go to the the main display page
- Looks like your tree has only numbers as leaf IDs. If these are NCBI species taxonomy IDs, we can try to automa
- interactively define color ranges for various internal nodes in your tree

Citations: Letunic and Bork (2006) *Bioinformatics* **23(1):**127-8 **and** Letunic and Bork (2011) *Nucleic Acids Res* **doi: 10.1093/nar/gkr201**
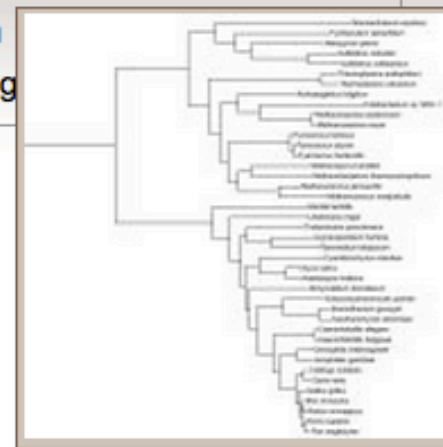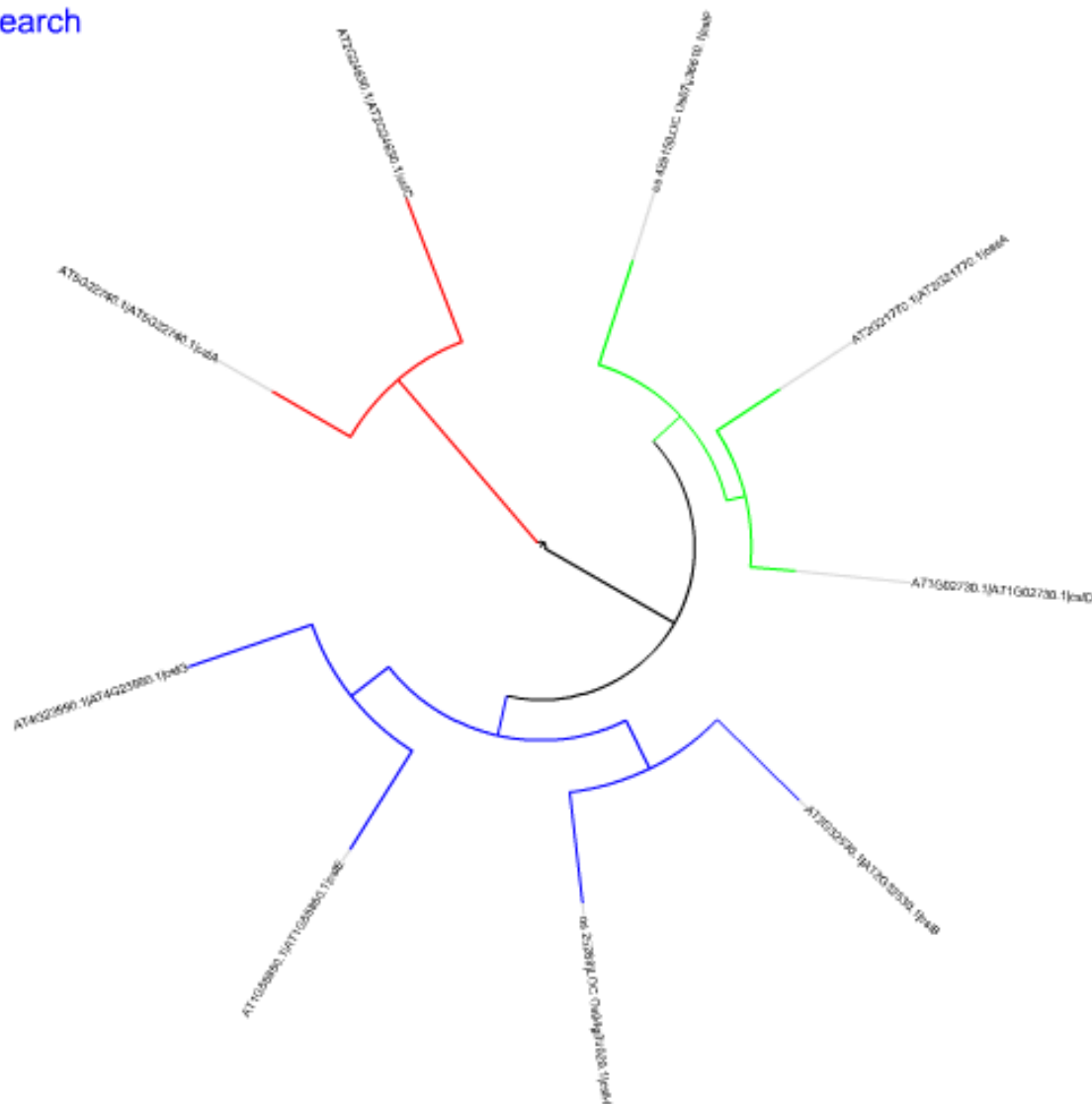
Tree: 131156211134574...   Views: current | saved

Basic controls

Advanced controls

Display controls

Mode: Circular | Circular mode | Normal mode | Unrooted mode

☐ Ignore bran...

Update tree

Reset tree   Export tree

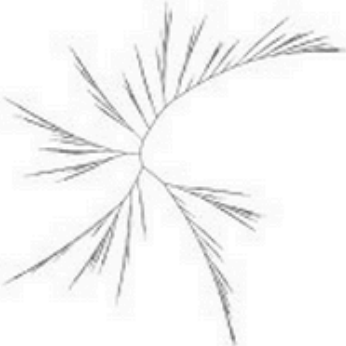⊕ Tree scale: 0.1 ⊢

Search

⊖

More options to display the phylogram

Export the tree
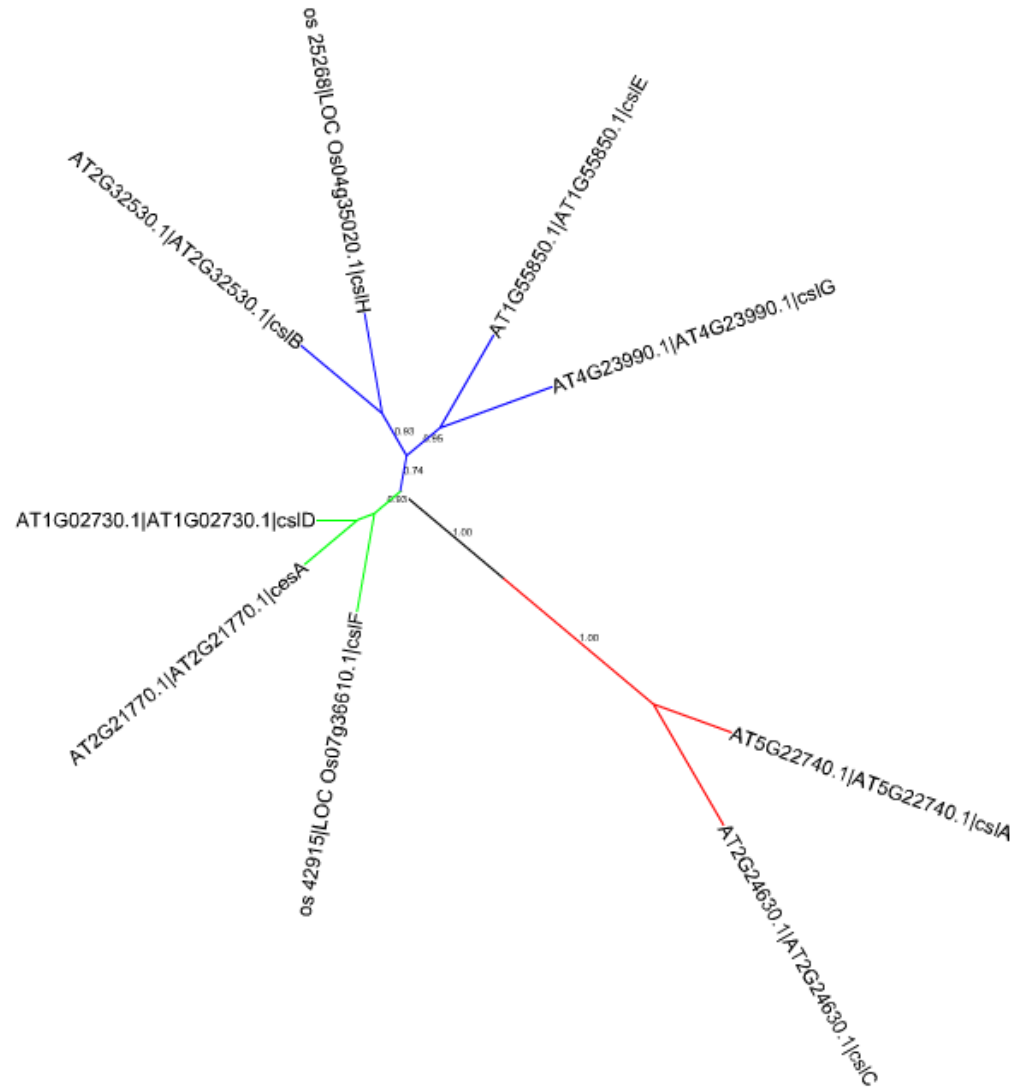
Export current tree (*131156211134574136200029650*) to other formats

Current display mode:	Other tree features:	
Unrooted	**Leaves visible:**	9
	Rotation:	0°
	Arc:	360°
	Branch lengths ignored:	No
	Branch lengths displayed:	No
	Bootstraps displayed:	Yes, >0.7 as text

Please select the output format below and customize other export op

Format	Portable Document Format (pdf) ⬍	select t...
Font size	30 pixels	font size
Display leaf labels	● yes ○ no	you can
Line width	1 px	line wid
Branch colors	○ ignore ● use	if branc
Leaf labels	○ Show original ID ● Show current label	if leaf labels were edited (or automatically assigned

Bitmap formats
Portable Network Graphics (png)
Vector formats
Encapsulated Postscript (eps)
Postscript (ps)
✓ **Portable Document Format (pdf)**
Scalable Vector Graphics (svg)
Text formats
Newick (txt)
Nexus (txt)
PhyloXML (txt)

Select the dataset(s) you want displayed:

Note: There are no restrictions on the number of datasets you can include in the exported trees. iTOL will not chec during post processing.

Dataset types which cannot be displayed in the currently selected mode have been disabled.

☐ ▇ **color** (color gradient)

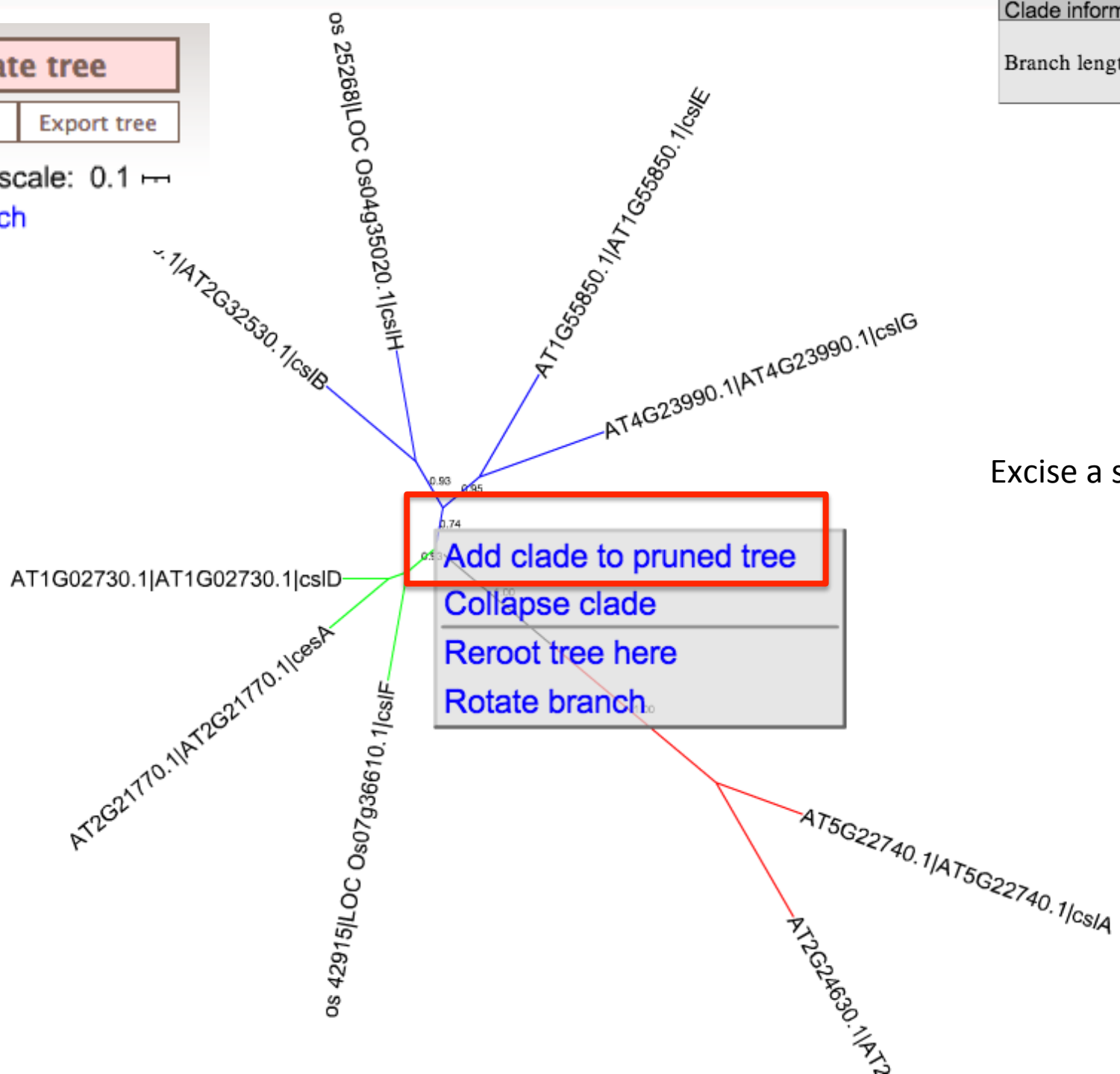● Do not include any datasets

Export tree

Excise a subtree